A fast and accurate Data-Driven Model for estimating the production temperature of High-Temperature Aquifer Thermal Energy Storage

Geerts, David; Daniilidis, Alexandros; Liu, Wen

DOI
10.1016/j.applthermaleng.2025.126817

Publication date
2025

Document Version
Final published version

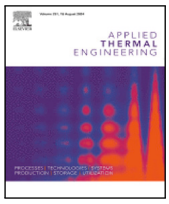
Published in
Applied Thermal Engineering

Applied Thermal Engineering 278 (2025) 126817

Contents lists available at ScienceDirect

Applied Thermal Engineering

journal homepage: www.elsevier.com/locate/ate



Research Paper

A fast and accurate Data-Driven Model for estimating the production temperature of High-Temperature Aquifer Thermal Energy Storage

David Geerts [a,*], Alexandros Daniilidis [b], Wen Liu [a]

[a] Copernicus Institute of Sustainable Development, Heidelberglaan 8, 3584 CS, Utrecht, The Netherlands
[b] Faculty of Civil Engineering and Geosciences, Stevinweg 1, 2628 CN, Delft, The Netherlands

ARTICLE INFO

Dataset link: https://github.com/dayfix/DD_ATES

Keywords:
Data-driven model
High-Temperature Aquifer Thermal Energy Storage
Nearest neighbor search
Recovery efficiency prediction

ABSTRACT

High-Temperature Aquifer Thermal Energy Storage (HT-ATES) has the potential to significantly increase the renewable heat share in heating systems. However, HT-ATES has not been implemented in the current energy system models because the widely applied numerical models for HT-ATES are computationally expensive. This leads to a lack of HT-ATES assessment from an energy system perspective. Therefore, an accurate and computationally efficient model that is widely applicable is needed to facilitate such implementation. This research aimed to develop a novel data-driven model that generates the temperature profile of an HT-ATES accurately and computationally efficiently. A trained machine learning algorithm predicts the recovery efficiency for an HT-ATES system, which, combined with other parameters, enables a nearest neighbor search to identify a suitable temperature profile. As a result, the temperature profile generated by the data-driven model has a root mean square error of 1.22 °C compared to the numerical model output. This error was shown to be larger for lower recovery efficiency values compared to higher values. The machine learning algorithm used to predict the recovery efficiency has a root mean square error of 1.45 percentage points. The data-driven model has a computation time of less than half a second, which is more than 180,000 times faster than the numerical model that was used to generate the data. This model is, therefore, suitable for integration in larger energy system models.

1. Introduction

There are growing concerns about CO_2 emissions and their impact on global climate change. The heating sector is responsible for 40% of our global energy consumption [1] and is an important sector for reducing CO_2 emissions. High-Temperature Aquifer Thermal Energy Storage (HT-ATES) systems offer a promising solution to reduce CO_2 emissions, as they store excess heat, which can be used instead of burning fossil fuels when demand is high in winter [2]. HT-ATES is especially relevant in combination with less flexible sustainable heat sources such as solar and geothermal, as it can increase the load of these sources by shifting the use of the produced heat to the colder season, therefore also reducing CO_2 emissions [3].

The feasibility of an HT-ATES system is highly dependent on the efficiency of the system, which is determined by the temperature of the water extracted from the HT-ATES, also called the temperature profile. There are different methods to calculate the temperature profile. Most common are the numerical models that simulate the physics of fluid flow in the aquifer, of which many different methods exist. Examples are finite differences method [4,5], finite element method [6–8] and finite volume method [9]. In [10], 11 different simulators were compared using the mentioned methods and showed that the simulation results are often comparable and accurate. However, these models are reported to have a high computational load, leading to long run times for individual assessments between 1 and 10 h [11–14]. The long computational time and complexity hinder their application in larger energy system modeling. As shown by Lyden et al. [15], no energy system modeling tool has yet implemented an HT-ATES model, resulting in a significant gap in the analysis of HT-ATES from an energy system perspective. This highlights the critical importance of developing a computationally efficient and accurate model that can be integrated into energy system tools, enabling a large number of simulations to be conducted efficiently. By reducing computational demands, such a model would facilitate comprehensive scenario analyses, sensitivity studies, and optimization tasks in a time-efficient manner.

Another method for calculating the temperature profile is the analytical method. Only one analytical solution for estimating the temperature profile has been suggested [16]. They proposed an analytical approach to derive the efficiency and the corresponding temperature

* Corresponding author.
E-mail address: d.c.geerts@uu.nl (D. Geerts).

https://doi.org/10.1016/j.applthermaleng.2025.126817
Received 27 January 2025; Received in revised form 7 April 2025; Accepted 12 May 2025
Available online 27 June 2025

Nomenclature

η	Recovery efficiency (–)
η_{data}	Recovery efficiency in the dataset (–)
η_{pred}	Recovery efficiency predicted by the ML algorithm (–)
$\mathbb{P}^{data}$	Value of a parameter in the data set (–)
$\mathbb{P}^{n}$	New set of parameters inputted in the data-driven model (–)
$\mathbb{T}_{data}$	Temperature profile over time of the dataset (°C)
$\mathbb{V}_{data}$	Injection pattern over time of the dataset (°C)
$\mathbb{X}$	Temperature profile over volume (°C)
$\overline{T}_i$	Average temperature of injected water (°C)
$\overline{T}_o$	Average temperature of extracted water (°C)
σ	Deviation in temperature profile (°C)
a	Anisotropy (–)
d	Relative distance between two data points (–)
E_{in}	Energy injected into of well (J)
E_{out}	Energy extracted out of well (J)
H	Thickness aquifer (m)
k_h	Horizontal hydraulic conductivity (m day^{-1})
k_v	Vertical hydraulic conductivity (m day^{-1})
n	Porosity of aquifer (–)
T_e	Temperature of extracted water (°C)
T_g	Ambient groundwater temperature (°C)
T_i	Temperature of injected water (°C)
V_e	Yearly extracted water volume (m^3)
V_i	Yearly injected water volume (m^3)
$T_i^{\mathbb{P}^n}$	Temperature of injected water in $\mathbb{P}^n$ (°C)
DDM	Data-Driven model (–)
HT-ATES	High-Temperature Aquifer Thermal Energy Storage (–)
ML	Machine Learning (–)
RMSE	Root Mean Square Error (–)

profile, which would resolve the computational burden that numerical models have. However, the scope in which this analytical approach is accurate is narrow, which was shown by the fact that when changing either injection rate, diffusivity, or injection screen length, the accuracy decreased, limiting the applicability of this method.

Based on the discussion above, a model is needed to calculate the temperature profile of HT-ATES systems that can reduce computational load while preserving the accuracy of numerical models across a wide range of HT-ATES parameters. Such a model facilitates the integration of an HT-ATES model into an energy system modeling tool and would support the economic and environmental assessment of HT-ATES from an energy system perspective. To address this need, we develop a Data-Driven Model (DDM) enabling computationally efficient and accurate temperature profile predictions. The following definition of temperature profile is used in this study: the temperature of the volume extracted from a well over time. Note that this definition of the temperature profile inherently depends on the rate of extraction.

The DDM was created based on a large dataset previously generated [11], which contains different parameter values and the corresponding temperature profiles. An accurate data point should be found in this dataset. The temperature profile is reflected in three aspects: the maximum and minimum temperature reached during the extraction period and the path between those points. The maximum and minimum temperatures reached during extraction are approximated using the temperature of the injected water and the ground temperature, respectively. The path between the maximum and minimum is reflected by the Recovery Efficiency (η) of that cycle. The η is often calculated using the mentioned numerical models. To avoid reliance on these models, the η is instead predicted by a Machine Learning (ML) algorithm, which is well-suited to accurately predict single values, in this case, the η [17], while being computationally efficient. Using these three parameters, which reflect important aspects of a temperature profile, a nearest-neighbor search is conducted to find the most accurate temperature profile in the dataset. This search ensures that the resulting temperature profile is accurate while the approach is computationally efficient.

Based on the above discussion, the key novelties of this work are the following: (1) DDM methodology: By integrating ML with a nearest-neighbor search, the DDM effectively identifies accurate temperature profiles. This approach combines the predictive accuracy of ML with the physical constraints inherent to temperature profiles, enforced through a search algorithm that is constrained by full-physics numerical models, ensuring consistency with comprehensive numerical model results. (2) Wide applicability: The dataset's extensive parameter range allows the DDM to be applied to a wide variety of HT-ATES systems. Lastly, (3) Computational efficiency and accuracy: the proposed model is based on data and relatively simple equations compared to numerical modeling, resulting in a low computational cost, which, combined with the wide range of applicability, makes this model useful for a large range of applications.

This paper describes the DDM and its applicability, accuracy, and limits. First, the DDM is explained in more detail in Section 2, after which Section 3 explains the assessment method for the model. The results of the model assessment are presented in Section 4 and discussed in Section 5. Conclusions are drawn in Section 6

2. Data-driven HT-ATES model

This section outlines the DDM used to generate a temperature profile. Firstly, Section 2.1 introduces the dataset and describes its content. Section 2.2 explains how ML is applied to predict the η. Section 2.3 describes how this η is used in combination with the nearest neighbor search in the dataset to find the closest matching temperature profile. Lastly, Section 2.4 explains how the closest matching temperature profile is adapted to better align with the used parameters. Section 3 explains the testing of this model as well as the assessment of the boundaries. The structure of this section and the next section is visually explained in Fig. 1. The model is open source and can be found on Github [18].

2.1. Dataset description

The data used was obtained from the research in Geerts et al. (2024) [11]. In that paper, multiple simulations were done to demonstrate the relation between the η of an HT-ATES system and seven relevant parameters. Where the η is defined by Bloemendal et al. [4] as

$$\eta = \frac{E_{out}}{E_{in}} = \frac{V_e \Delta T_e}{V_i \Delta T_i} = \frac{\overline{T_e} - T_g}{\overline{T_i} - T_g}. \tag{1}$$

The volume injected is assumed to be the same as the extracted volume, allowing for an unambiguous comparison of η values. The seven parameters and their minimum and maximum values are shown in Table 1, where anisotropy is defined as k_h/k_v.

The dataset contains 3501 data points, each point contains the mentioned seven parameters' value, the corresponding η, and the corresponding temperature profile for the first 8 years of operation. An example of a small part of the dataset is shown in Table 2. The DDM is given a new set of parameter values, called $\mathbb{P}^n$, and generates a new temperature profile based on these parameters.

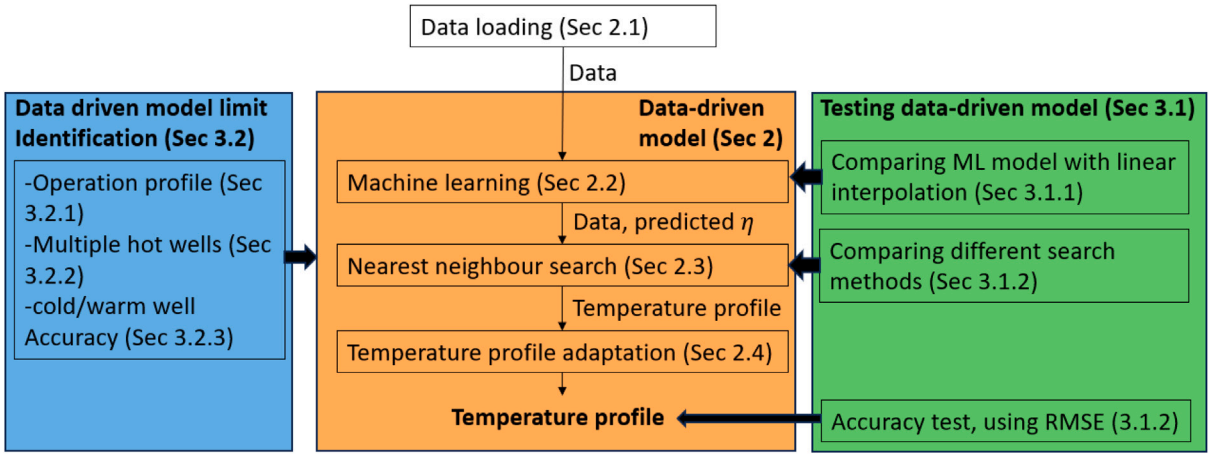


Fig. 1. The structure of Section 2 and 3 and the relation between the different sections.

Table 1
Minimum and maximum value of the parameters in the dataset [11].

Parameter	Minimum value	Maximum value	Unit
Porosity	0.1	0.3	–
Yearly injected volume	10^4	10^6	m^3
Injected temperature	25	80	°C
Ambient ground temperature	10	30	°C
Horizontal hydraulic conductivity	1	85	m day^{-1}
Anisotropy	1	100	–
Aquifer thickness	20	104	m

2.2. ML algorithm

Next, the η was predicted based on $\mathbb{P}^n$. The η generally increases during the first few years of operation [6,19], after which the η stabilizes when a dynamic equilibrium is reached with the aquifer, which means the η does not change in subsequent years. The predicted η is the stabilized η that was calculated after eight injection and extraction cycles [11]. The ML algorithm predicts the η based on the mentioned seven parameters. It was trained and tested on the dataset described in the previous section using 80% of the data for training the algorithm and the other 20% for testing [20,21]. This split made sure that the resulting accuracy of the ML model was not caused by overfitting. The applied ML algorithm is an extreme gradient boosting regression [22] as is implemented in the XGboost Python package [22]. XGboost was shown to be accurate by multiple authors in multiple fields, such as global solar irradiance prediction and district heating load forecasting [17,20,21,23]. This ML algorithm was then used within the DDM to predict the η based on $\mathbb{P}^n$.

2.3. Nearest neighbor search

The resulting predicted η was used to find the closest matching temperature profile in the dataset. Reformulating Eq. (1) to solve for recovered temperature yields:

$$\overline{T_e} = \eta(\overline{T_i} - T_g) + T_g, \tag{2}$$

shows that the average extracted temperature is dependent on the η, injected temperature, and ambient groundwater temperature. This average extracted temperature is dictated by the temperature profile, but the temperature profile cannot be calculated when only the average extracted temperature is known. Therefore, the three variables on the right side of the formula were used to find the closest temperature profile in the dataset, using a nearest neighbor search approach.

This approach calculates the relative distance between $\mathbb{P}^n$ and the parameters in the dataset, only taking into account the three parameters shown on the right side of Eq. (2). This distance (d) between a data point and $\mathbb{P}^n$ was calculated as follows

$$d = \sqrt{\sum_{i=1}^{N_{para}} \frac{P_i^{data} - P_i^n}{P_i^{max} - P_i^{min}}}, \tag{3}$$

where N_{para} is the number of parameters taken into account, which is three this case (see Eq. (2)). P_i^{data} refers to the value of the data point of parameter i and P_i^n refers to the new value inputted in $\mathbb{P}^n$ for that same parameter. P_i^{max} and P_i^{min} refer respectively to the maximum and minimum value of parameter i as shown in Table 1. For η, the maximum value is 1, and the minimum value is 0. This equation calculates the relative distance between a data point and $\mathbb{P}^n$, where each parameter is taken into account equally when determining the distance. This distance is calculated for all 3501 data points described in Section 2.1, and the temperature profile of the data point with the lowest d value is used in the next step.

2.4. Temperature profile adaptation

The adaptation was to correct for the temperature of the injected water (T_i). The temperature of the injected water of the data point chosen in the last step can be higher or lower than the injected temperature in $\mathbb{P}^n$, leading to inaccuracy. The maximum temperature that the temperature profile should reach should be equal to the temperature of the injected water (T_i) in $\mathbb{P}^n$ ($T_i^{\mathbb{P}^n}$). The temperature profile was not corrected for the ground temperature (T_g) even though the T_g in $\mathbb{P}^n$ might be different from the T_g of the data point. This is because T_g is only an approximation of the minimum temperature reached and not a strict minimum that the temperature profile should reach. This injected temperature was corrected by using the following equation

$$\mathbb{T}_{man} = \mathbb{T}_{data} * \frac{T_i^{\mathbb{P}^n}}{max(\mathbb{T}_{data})}. \tag{4}$$

Here max refers to taking the maximum value in the temperature profile. This adapts the temperature profile to better suit the new injected water temperature. This $\mathbb{T}_{man}$ is the output of the model, which is the temperature profile over time for the first 8 years of operation. For temperature profiles beyond the eighth year, the temperature profile of the eighth year can be used as the η is stabilized and should not change in subsequent years.

As mentioned in the introduction, the temperature profile always implicitly depends on the extraction rate. To make this dependency explicit, the temperature profile over volume is introduced to show the temperature of the extracted water over volume instead of over

Table 2
Example of the layout of the dataset used for the DDM, where the temperature profile contains a time series.

Index	n	V_i	T_i	T_g	H	k_h	a	η	Temperature profile
0	0.1	1e6	80	30	20	1	1	0.83	[80, 80 …, 80, 80]
…	…	…	…	…	…	…	…	…	…
3501	0.3	1e6	25	10	105	85	100	0.93	[25, 25, …, 25, 25]

time. This decouples the production temperature from the operation profile, where the operation profile is defined as both the injection and extraction rate of one well. Temperature over volume allows the model to be applied to operation profiles beyond the ones considered in the dataset (which is the Base case in Fig. 2), adding great flexibility to the model. This introduces some error because heat losses in the aquifer are also proportional to time, which is not explicit in the temperature profile over volume [16]. The accuracy of this approach will be discussed in Section 4.

The temperature over volume was obtained using the following formula

$$\mathbb{X} = \frac{\mathbb{T}_{man}}{\mathbb{V}_{data}}. \tag{5}$$

This equation divides the temperature profile over time ($\mathbb{T}_{man}$) by the used operation profile over time ($\mathbb{V}_{data}$), resulting in the temperature profile over volume ($\mathbb{X}$).

3. DDM assessment and limit identification

First, in Section 3.1, the testing of the accuracy of the DDM is explained. Next, in Section 3.2, the method of assessing the limits of the DDM is explained.

3.1. Accuracy assessment of DDM

3.1.1. ML algorithm accuracy assessment method

The ML algorithm explained in Section 2.2 was tested. The predictive performance of the ML algorithm was captured using the Root Mean Square Error (RMSE) as follows [24]

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\eta_{pred} - \eta_{data})^2}{n}}, \tag{6}$$

where n is the number of observations in the testing dataset. This formula shows the error in the predictions, where larger errors have a larger impact on the RMSE. The ML algorithm was compared with a linear interpolation algorithm. The linear interpolation algorithm interpolates between the η values of the dataset based on all parameter values of the tested data point and uses the scipy.interpolate.interpn function [25].

3.1.2. Nearest neighbor accuracy assessment method

The distance (calculated in Eq. (3)) was based on three parameters, for reasons explained in Section 2.3. However, this distance can also be calculated using any combination of parameters. To check the accuracy of only using the three parameters mentioned in Section 2.3, multiple combinations of parameters were compared. To facilitate this comparison, five options were created, and each calculates the distance of Eq. (3) based on different parameters. The five options and the used parameters are shown in Table 3.

These options were compared using the following method. First, the parameters of a data point were obtained. This data point was then removed from the dataset to prevent the nearest neighbor algorithm from finding that data point. The multiple nearest neighbor options were run, generating a temperature profile. This temperature profile was compared with the temperature profile of the removed data point, using the RMSE. This process was repeated for every data point in this way, and the DDM is directly compared with the results of the numerical model. This method of calculating the RMSE was also used to obtain the RMSE of the DDM in general.

Table 3
Options of parameters included in distance calculation of the nearest neighbor search, used for comparison.

Options	Included parameters
Base	η, T_i, T_g
η only	η
All parameters	η, T_i, T_g, V_i, a, k_h, H, n
Four random parameters	a, T_i, H, V_i
Base plus one	η, T_i, T_g, V_i

3.2. Limits identification

The original dataset was based on a numerical model that included certain assumptions. When these assumptions are altered, the accuracy of the DDM may decrease, limiting the reliable use of the DDM. One operational assumption is tested, namely, using a different operation profile than the one used to create the data points. The test aims to evaluate how changes in these operational assumptions affect the accuracy of the data-driven model, which is done by changing the assumption in the MODFLOW model [11] and comparing the resulting temperature profile with the temperature profile in the dataset.

The yearly operation profile was assumed to be a sinusoid as shown in the base case in Fig. 2. This is called base, as it is the operation profile on which the dataset is based and which mimics the yearly variability in heat supply and demand. The effect of keeping the operation profile the same between simulations is twofold: firstly, the injection and extraction period are sequential, with each period lasting six months. Secondly, the injected and extracted volumes always follow the same pattern. The consequence of these assumptions is tested by running the model from [11] again using different operation profiles.

Three operation profiles were created to test the effect of keeping the operation profile the same between simulations: (1) sequential random operation pattern, where sequential means that the extraction and injection period are sequential, lasting six months each, and random means that the injected and extracted amount is randomized for each time step. (2) sequential constant pattern, where constant refers to the fact that the injected and extracted amount are kept constant during the extraction and injection period, and (3) mixed random operation profile, where the extraction and injection period are shorter than six months. This was implemented as three months of injecting, three months with both injecting and extracting, switching between injecting and extracting multiple times at random intervals, three months of extracting, and again three months with both injecting and extracting at random intervals. The used operation profiles are shown in Fig. 2. To allow for a fair comparison, the extracted and injected volume over a year was kept the same between operation profiles, which was the assumption used in Eq. (1).

The temperature profiles of these operation profiles are compared with the temperature profile of the base profile, which is the operation profile used for generating the 3501 data points (shown in Fig. 2 as base). The goal of this comparison is to identify which operational profiles can be accurately predicted by the DDM. The deviation of the different operation profiles from the base profile is calculated as follows:

$$\sigma = \frac{\sum_{i=1}^{N_{data}} |T_{base,i} - T_{scen,i}|}{N_{data}} \tag{7}$$

where σ is the deviation and $T_{base,i}$ and $T_{scen,i}$ are the temperature value at point i for the base operation profile and the operation profile it

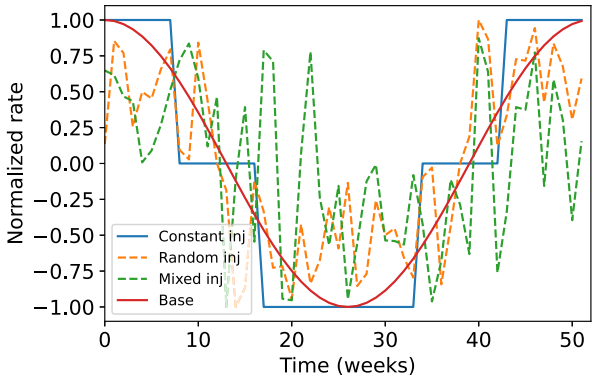


Fig. 2. Different operation profile used for comparison, positive rate refers to injecting and negative rate refers to extracting. The total volume used is the same between operation profiles.

is compared with respectively. N_{data} is the number of data points in one temperature profile. This σ is calculated for 100 data points using different parameter values, using only parameter values corresponding to a minimum or maximum value as shown in Table 1, to test the edge cases.

4. Results

4.1. Model accuracy results

4.1.1. Machine learning algorithm

Predictions from the ML algorithm were more accurate than those achieved by linear interpolation (Fig. 3). The RMSE was 1.45 percentage points for ML compared to 16 percentage points for linear interpolation. The linear interpolation was accurate for those data points that were very close to any of the points that the algorithm interpolates between. The algorithm calculates η values similar to those of the points interpolated between, and if the data point is close to one of these points, the η value generated by linear interpolation is very similar, which is accurate. However, for other points, this was inaccurate, showing that there is no linear relation between the parameters and the η. The ML algorithm was, on average, better able to predict the η values. The ML model was

4.1.2. Nearest neighbor options comparison

The RMSE of the nearest neighbor options were compared (Table 4). As can be seen, the "Base" option has the lowest RMSE; only the "Base plus one" option is comparable, while the other options perform significantly worse. As explained in Section 2.3, the calculation of $\overline{T_e}$ is based on these three parameters, and therefore, the temperature profile can best be found using only these three parameters. The "Base" option is generally the most accurate, although it does not perform best for all tested data. Most notable is that the "Base" option does not perform well with a small injected volume. The "Base plus one" option performs better, as it includes injected volume in its search.

An example of temperature profiles of the different options is shown in Fig. 4. The Base option has the lowest deviation. As can be seen, all options have taken a different data point on which the temperature profile is based because all options have a different temperature profile. If two options had chosen the same data point, then their resulting temperature profile would be identical. Some of the data points have a more suitable temperature profile than others, leading to a lower deviation.

Another design for the nearest neighbor search was also considered. This design used different weights for each parameter when using Eq. (3). Using different weights changes the model to be more accurate for some data points and less accurate for others, but no weights were found that led to a significant reduction in RMSE.

Table 4
Comparison RMSE of the different nearest neighbor options, calculated using all data points.

	Base	η only	All parameters	Four random	Base plus one
RMSE	1.22 °C	3.48 °C	3.47 °C	3.79 °C	1.40 °C

4.1.3. Computational performance

The RMSE of the DDM is 1.22 °C compared to the numerical model that used the same parameters and can thus be considered the reference. The run time of the model was on average 0.20 s per temperature profile and 28% of this was spent on loading the data set, which used 235 megabyte of memory (Section 2.1), 6% on predicting the η (Section 2.2), another 47% on the nearest neighbor algorithm (Section 2.3) and the last 19% on the temperature profile adaptation (Section 2.4). This was timed using Python 3.9 and an Intel Core i7-1255U without multiprocessing. The numerical model on which this DDM is based required, on average, 601 min for one temperature profile, making this model 180,000 times faster.

4.1.4. Temperature profile validation

In Fig. 5 two examples of a temperature profile generated by the DDM are shown, one that has the highest deviation in the dataset (Fig. 5(a)) and one which has a deviation equal to the RMSE (Fig. 5(b)). Only 10% of the dataset has a higher deviation than 1.22 °C. The large deviation is caused by the difference in the rate of decline of the temperature profile. Both temperature profiles have a comparable η. However, the temperature profile of the DDM model drops at the beginning of the extraction phase and then stabilizes, while for the data point, the extraction temperature decreases gradually, leading to a large deviation.

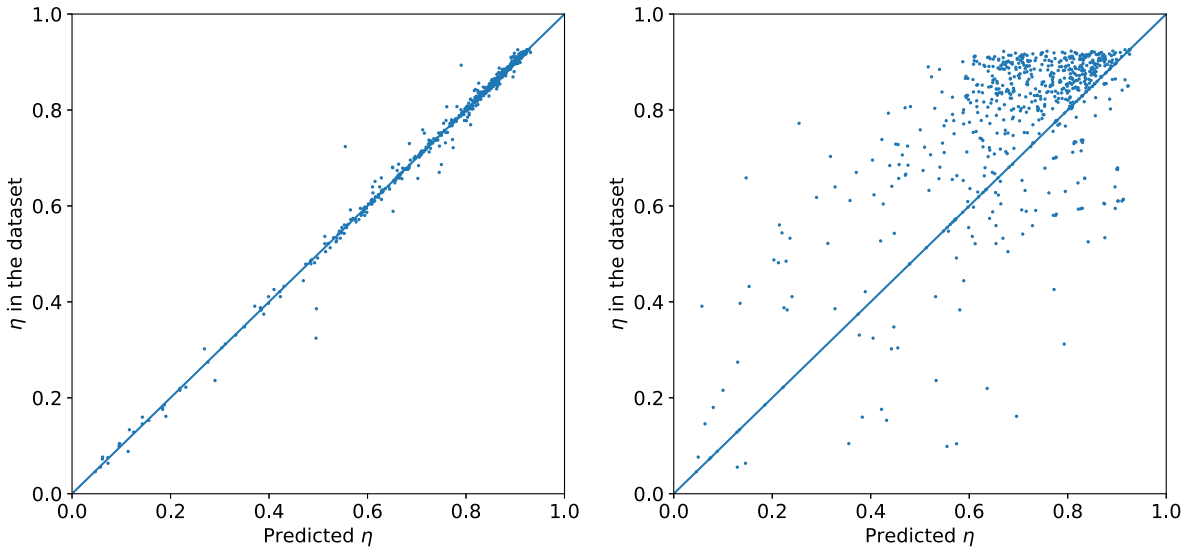
The temperature profile with a deviation equal to the RMSE (Fig. 5(b)) still looks very similar to the temperature profile of the dataset, and a large part of the deviation can be attributed to the first extraction phase. Where the temperature profile of the data point reaches a very low temperature compared to the temperature profile of the DDM. However, when only using the eighth year, the error is only 0.4 °C.

The error in the temperature profile generally increases when the η decreases (Fig. 6). This is likely because fewer data points have a low efficiency (Fig. 6). When using the nearest neighbor search with a low η, the closest temperature profile is likely less accurate as there are fewer points to choose from. Therefore, with lower η, the error is likely higher, but the technical potential of HT-ATES systems with low η is also low.

4.2. Identifying model limits

The η values of the different operation profiles shown in Fig. 2 were analyzed. The η of the sequential constant and sequential random operation profile differed on average by 0.5% and 0.3%, respectively, from the base case. Where the η value of the mixed operation profile was on average 9.2% higher than the base scenario. This is because using a mixed operation profile reduces the time between storage and extraction, leading to reduced storage periods, which in turn leads to reduced losses to the surroundings, increasing the η. The other two operation profiles are similar to the base operation profile in terms of time between storage and extraction, leading to similar η values.

The parameter values used for Figs. 7(a) and 7(b) are shown in Table 5, and the η of the base case operation profile is 88%. Fig. 7(a) shows the temperature profiles of the different operation profiles. What is immediately clear is that the mixed profile shows peaks, in contrast to the other temperature profiles, which are only decreasing during the extraction period. These peaks correspond with injecting instead of extracting. After injection, the temperature out of the hot well is the same as the temperature injected, which corresponds with these



(a) Predictions using ML algorithm based on the test set

(b) Predictions using linear interpolation based on the test set

Fig. 3. Comparison ML algorithm against linear interpolation, showing the error in the predictions.

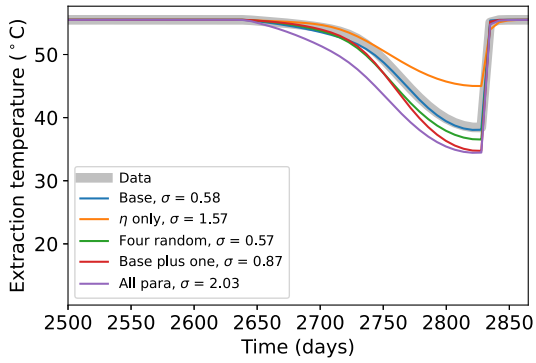


Fig. 4. Temperature profile comparison for the different nearest neighbor options for the eighth year.

peaks. For the other two operation profiles, the temperature profiles are comparable to the base case, but the injection and extraction periods differ, meaning that at the same time step, each operation profile can have extracted more or less volume than the base case, leading to a different temperature at that time step. This is most notable with the constant operation profile, which injects at different time intervals than the base case, and large differences can be seen during these time intervals.

However, the temperature over volume ($\mathbb{X}$) can resolve these time differences (Fig. 7(b)). The sequential operation profile outputs are in better agreement with the base operation profile. For the mixed operation profile, the temperature profile over volume still has large errors, and the conclusion needs to be drawn that this model cannot predict the temperature output of a mixed operation profile. The model underestimates the performance of an HT-ATES that has a mixed operation profile, resulting in η on average 9.2% higher than the base case scenario.

Table 6 shows the deviation of the random and constant operation profile. The temperature over volume deviation is significantly lower than the temperature profile deviation for both operation profiles, showing that this $\mathbb{X}$ is suitable when the operation profile differs from the base case.

Table 5
The parameter values used in Figs. 7(a) and 7(b).

	n	V_i	T_i	T_g	H	k_h	a
Value	0.3	1E6	25	10	20	85	1

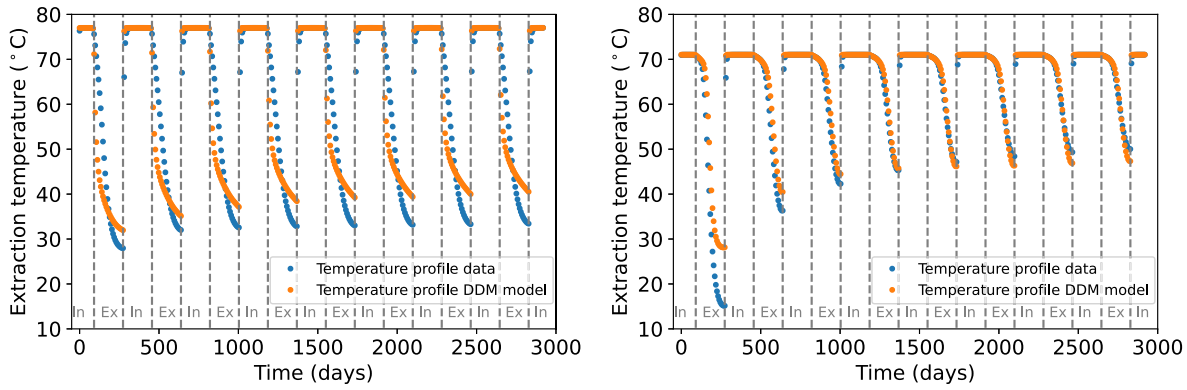
Table 6
Average deviation (σ) of the two injection patterns.

Injection pattern	Constant	Random
Temperature profile	1.62 °C	0.53 °C
Temperature over volume	0.17 °C	0.40 °C

4.2.1. Shifting injection profile

An inconsistency in the resulting temperature profile was found relating to the first year of operation, which contained one injection and one extraction period. As seen in Fig. 2, the hot well starts with injecting but only injects half of the total yearly volume in the first storage cycle, from week 0 to week 13. After this, the full yearly volume is extracted. This leads to the first year of the temperature profile reaching a very low minimum compared to when the full total yearly volume would be injected and extracted.

As illustrated in Fig. 8, the temperature profile of the first year of the "Half total yearly injected volume before extraction" line reaches ground temperature at the end of the first year of extraction. This suggests that all the heat injected during the first injection cycle is extracted, resulting in minimal residual heat in the aquifer, and no increase in η is expected in the following year. Consequently, in the second year, the HT-ATES has not yet been heated, and the η for the second year does not benefit from the previous year. When injecting the full yearly volume before extraction, the minimum temperature reached during the first year's extraction aligns with the original operation profile for its second year, as depicted in Fig. 8. Therefore, if the HT-ATES starts with injecting the full total yearly volume before extraction, the first year of the temperature profile that is outputted by the DDM should be removed. This only influences the first year of operation, while the stabilized operation after eight years remains unchanged.



(a) Temperature profile with the highest deviation in the dataset, which is 8.3 °C

(b) Temperature profile with deviation equal to the RMSE of 1.22 °C

Fig. 5. Example of temperature profiles of the DDM against the temperature profile of the data set. The dashed lines refer to the switching of operation phases, with 'In' referring to injection and 'Ex' referring to extraction.

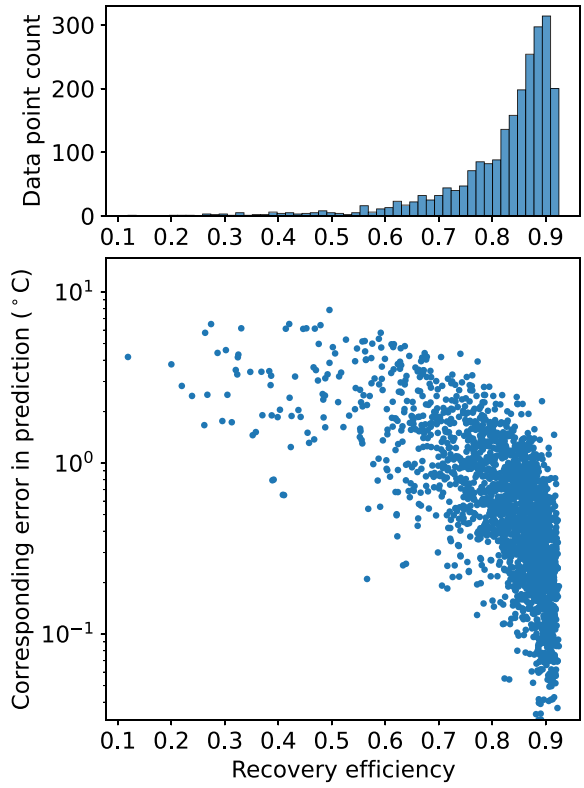


Fig. 6. Error of the individual data points against the η of the data point. With the top plot showing a distribution of η values in the dataset. Note the logarithmic scale on the bottom y-axis.

5. Discussion

The temperature profile generated by the DDM has an RMSE of 1.22 °C. The prediction of the η by ML algorithm has a RMSE of 1.45 percentage points compared to the numerical model. The ML algorithm would likely perform better with more data points, reducing the RMSE further. More data points also lead to higher chances of finding a more accurate data point during the nearest neighbor search. However, data generation was time intensive and took on average 601 min per data point. This model is accurate for data points with a high η, which are the most relevant HT-ATES systems. Where lower η generally leads to lower accuracy. Generating more data points with a low η would likely improve the accuracy, however, these HT-ATES systems are less interesting to install.

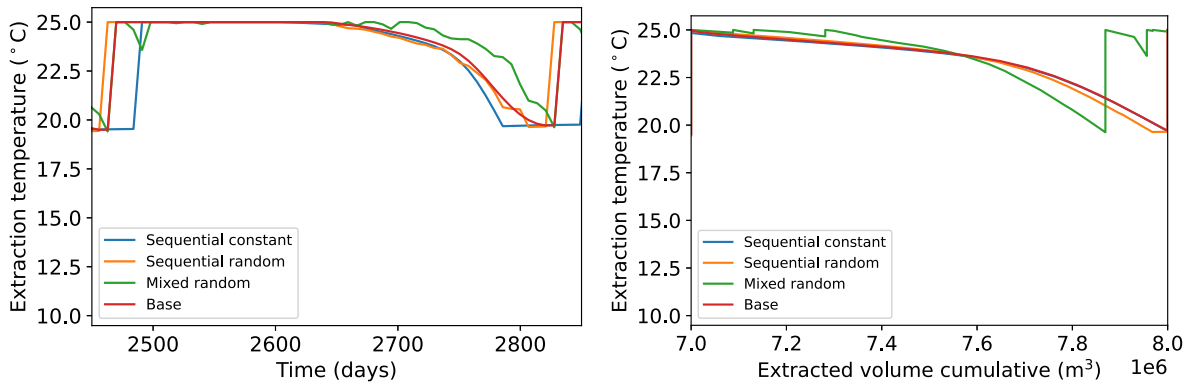
The error in temperature predictions can impact energy system design, especially in hybrid systems combining HT-ATES with other heat sources. With an RMSE of 1.22 °C, deviations in predicted temperatures can affect storage dispatch, HT-ATES sizing, and overall efficiency. Overestimation may lead to increased reliance on backup heating, while underestimation could cause underutilization of stored heat and unnecessary over-sizing. However, the error is lower for higher recovery efficiencies, which are the most relevant cases for practical HT-ATES applications. Additionally, the purpose of this model is to provide a quick method to estimate the performance of the HT-ATES. Detailed models are still recommended for the actual implementation of the HT-ATES. These detailed models can take into account practical considerations such as variability in demand or supply or regulatory constraints.

This DDM can be implemented to represent the HT-ATES system without the computational effort required for running numerical models. This efficiency makes the DDM very appropriate as a component within a larger energy system model. This facilitates larger models to also include HT-ATES systems and enables them to adequately determine the impact of implementing such an HT-ATES system from a system perspective. However, the performance of the HT-ATES when using a mixed operation profile is underestimated, which needs to be taken into account when implementing this model into larger models.

Other designs for the DDM could be considered, which could improve computational time, accuracy, or both. For example, training a ML algorithm to directly predict the temperature profile, which skips the nearest neighbor search step. However, with the data-driven approach, the output will always be constrained to the numerical modeling output, leading to temperature profiles that will always be comparable to the ones calculated by a full numerical model. When directly applying a ML algorithm, this is not guaranteed, and temperature profiles might be illogical.

There are also limitations to the DDM. Firstly, the assumption in [11] was that there should be no interaction between wells; therefore, for this model to be reliable, wells should be placed sufficiently far apart to prevent other wells from influencing the temperature profile of the targeted well.

Secondly, the DDM and the results in this paper are constrained by the underlying data. There is no knowledge of how this model behaves when using values outside of the parameter space. It will search for the closest data point it can find and will give a temperature profile similar to that data point, which might be inaccurate depending on the distance between the new values and the parameter space.



(a) Temperature profiles ($\mathbb{T}_{man}$). (b) Temperature profiles over volume ($\mathbb{X}$).

Fig. 7. $\mathbb{T}_{man}$ and $\mathbb{X}$ for the different operation profiles for the eighth year of operation, with the parameter values shown in Table 5.

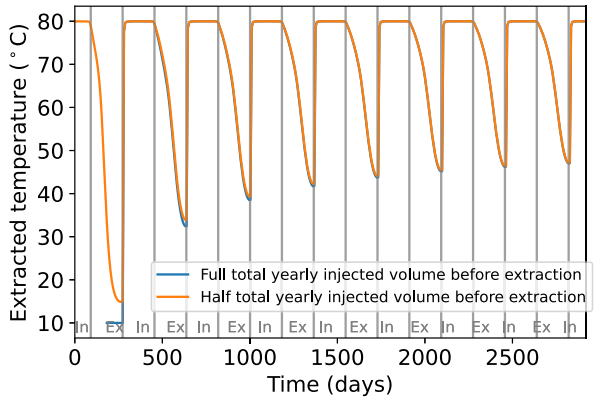


Fig. 8. Effect of start of operation on temperature profile, ground temperature = 10 °C and vertical lines reflect the change of operation mode. The full total yearly volume line starts in the second year to show the similarities with the half total yearly volume line.

The temperature profile created by the model is based on operation profiles that inject half of the total yearly volume before extracting the full total yearly volume, which is not always the case. The proposed solution is to skip the first year of operation because in the first year, the extracted heat is very close to the amount of injected heat. The temperature profile of the eighth year remains the same because this is the stabilized temperature profile, which is the same regardless of starting time.

With this model, future work could investigate the sizing of an HT-ATES system within the larger system, the optimization of flows within an energy system, or the comparison of different supply sources in combination with a storage component. Additionally, a more holistic approach of sizing heat supply technologies and HT-ATES simultaneously is enabled by this model. The design of this DDM can also be used for other purposes. It can be used to generate a time series for which data is available or can be generated, such as borehole thermal energy storage profiles [26] or the output of geothermal wells. This could also be a direction for future research.

6. Conclusion

This research contributes to the integration of High-Temperature Aquifer Thermal Energy Storage (HT-ATES) into larger energy system modeling tools by developing a novel data-driven HT-ATES model. The objective of the model is to generate an accurate temperature profile of a HT-ATES. The model is computationally efficient while maintaining the accuracy that numerical modeling provides. The model enables analysis of HT-ATES from a system perspective, facilitating its incorporation into broader energy system studies.

The Data-Driven Model required less than half a second to generate a temperature profile. In comparison, the MODFLOW model that was used to generate the data required an average of 601 min per data point, making the Data-Driven Model more than 180,000 times faster. The Root Mean Square Error of the model was shown to be 1.22 °C compared to the numerical model. The model can be used reliably with different operation profiles as long as the injection and extraction periods are sequential and last a few months each. However, it underestimates recovery efficiency when switching between injection and extraction within a month.

This study also found that for different operation profiles compared to the base case, temperature profiles over volume are more accurate than those over time, emphasizing the role of extracted volume in the temperature profile.

This research demonstrated both the usefulness and the limits of the Data-Driven Model. The data includes a wide range of HT-ATES parameters, making the model applicable across different HT-ATES systems. The model's robustness across varying operation profiles highlights its potential for broader applications in larger energy systems, maintaining accuracy while drastically reducing computational time.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded by the European Union under the Horizon Europe programme (grant no. 1011096566). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor CINEA can be held responsible for them.

Data availability

The Data-driven model and used data can be found here: https://github.com/dayfix/DD_ATES.

References

[1] EIA, Residential Energy Consumption Survey, Tech. Rep., US Energy Information Administation, 2023.

[2] P. Fleuchaus, S. Schüppler, M. Bloemendal, L. Guglielmetti, O. Opel, P. Blum, Risk analysis of high-temperature aquifer thermal energy storage (HT-ATES), Renew. Sustain. Energy Rev. 133 (2020) 110153.

[3] A. Daniilidis, J.E. Mindel, F. De Oliveira Filho, L. Guglielmetti, Techno-economic assessment and operational CO2 emissions of high-temperature aquifer thermal energy storage (HT-ATES) using demand-driven and subsurface-constrained dimensioning, Energy 249 (2022) 123682.

[4] M. Bloemendal, N. Hartog, Analysis of the impact of storage conditions on the thermal recovery efficiency of low-temperature ATES systems, Geothermics 71 (2018) 306–319.

[5] S. Beernink, N. Hartog, M. Bloemendal, M. van der Meer, ATES systems performance in practice: Analysis of operational data from ATES systems in the province of utrecht, The Netherlands, in: Proceedings of the European Geothermal Congress, 2019.

[6] H.A. Sheldon, A. Wilkins, C.P. Green, Recovery efficiency in high-temperature aquifer thermal energy storage systems, Geotherm. 96 (2021) 102173.

[7] C. Qi, R. Zhou, H. Zhan, Analysis of heat transfer in an aquifer thermal energy storage system: On the role of two-dimensional thermal conduction, Renew. Energy 217 (2023) 119156.

[8] L. Gao, J. Zhao, Q. An, X. Liu, Y. Du, Thermal performance of medium-to-high-temperature aquifer thermal energy storage systems, Appl. Therm. Eng. 146 (2019) 898–909.

[9] A. Réveillère, V. Hamm, H. Lesueur, E. Cordier, P. Goblet, Geothermal contribution to the energy mix of a heating network when using aquifer thermal energy storage: modeling and application to the Paris basin, Geotherm. 47 (2013) 69–79.

[10] J.E. Mindel, P. Alt-Epping, A.A. Les Landes, S. Beernink, D.T. Birdsell, M. Bloemendal, V. Hamm, S. Lopez, C. Maragna, C.M. Nielsen, et al., Benchmark study of simulators for thermo-hydraulic modelling of low enthalpy geothermal processes, Geotherm. 96 (2021) 102130.

[11] D. Geerts, A. Daniilidis, G.J. Kramer, M. Bloemendal, W. Liu, Analytically estimating the efficiency of high temperature aquifer thermal energy storage, Geothermal Energy 13 (1) (2025) 17.

[12] B. Bozkaya, R. Li, T. Labeodan, R. Kramer, W. Zeiler, Development and evaluation of a building integrated aquifer thermal storage model, Appl. Therm. Eng. 126 (2017) 620–629.

[13] M. Bloemendal, M. Jaxa-Rozen, T. Olsthoorn, Methods for planning of ATES systems, Appl. Energy 216 (2018) 534–557.

[14] P. Mugunthan, C.A. Shoemaker, Assessing the impacts of parameter uncertainty for computationally expensive groundwater models, Water Resour. Res. 42 (10) (2006).

[15] A. Lyden, C. Brown, I. Kolo, G. Falcone, D. Friedrich, Seasonal thermal energy storage in smart energy systems: District-level applications and modelling approaches, Renew. Sustain. Energy Rev. 167 (2022) 112760.

[16] D.W. Tang, H.H. Rijnaarts, Dimensionless thermal efficiency analysis for aquifer thermal energy storage, Water Resour. Res. 59 (11) (2023) e2023WR035797.

[17] J. Fan, X. Wang, L. Wu, H. Zhou, F. Zhang, X. Yu, X. Lu, Y. Xiang, Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China, Energy Convers. Manage. 164 (2018) 102–111.

[18] D. Geerts, DD ATES, 2024, https://github.com/dayfix/DD_ATES,

[19] S. Beernink, N. Hartog, P.J. Vardon, M. Bloemendal, Heat losses in ATES systems: The impact of processes, storage geometry and temperature, Geotherm. 117 (2024) 102889.

[20] P. Xue, Y. Jiang, Z. Zhou, X. Chen, X. Fang, J. Liu, Multi-step ahead forecasting of heat load in district heating systems using machine learning algorithms, Energy 188 (2019) 116085.

[21] C.N. Obiora, A. Ali, A.N. Hasan, Implementing extreme gradient boosting (xgboost) algorithm in predicting solar irradiance, in: 2021 IEEE PES/IAS PowerAfrica, IEEE, 2021, pp. 1–5.

[22] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, New York, NY, USA, ISBN: 978-1-4503-4232-2, 2016, pp. 785–794, [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785.

[23] R. Cai, S. Xie, B. Wang, R. Yang, D. Xu, Y. He, Wind speed forecasting based on extreme gradient boosting, IEEE Access 8 (2020) 175063–175069.

[24] T.O. Hodson, Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not, Geosci. Model. Dev. Discuss. 2022 (2022) 1–10.

[25] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, İ. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods 17 (2020) 261–272, http://dx.doi.org/10.1038/s41592-019-0686-2.

[26] A. Choudhary, R. Majumdar, S.K. Saha, Hybridisation of geothermal source with ORC-based load loop for uninterrupted generation of steady power, Int. J. Sustain. Energy 41 (1) (2022) 58–84.