PROCEEDINGS, 50th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 10-12, 2025
SGP-TR-229

Historical Data Prediction for Geothermal Systems Using Data-Driven Modelling

Daniel Clark, Michael Terekhin, Andreas Kempa-Liehr, John P. O'Sullivan, Michael J. O'Sullivan and Michael J. Gravatt

The University of Auckland, 70 Symonds Street, Grafton, Auckland, New Zealand, 1010

michael.gravatt@auckland.ac.nz

Keywords: Bayesian, Optimization, Production, Data-inference

ABSTRACT

Geothermal reservoir modelling requires detailed historical extraction data to predict a geothermal system's future state accurately. However, this data is rarely available to modellers. Often, wells are sparsely measured, while grouped measurements, such as mass flow at a separator, occur more frequently. Previously, this data history was manually estimated, leading to an inaccurate and subjective dataset with unquantifiable uncertainty. This is a serious problem for geothermal reservoir models relying on accurate, well-by-well data to generate predictions. This paper outlines the process undertaken to develop methods of predicting this unknown data history.

This paper introduces two data-driven models to address this issue. In the first method, is an optimal Tikhonov-regularized Linear Least Squares Optimization (TRLLSO) model, which is a computationally efficient and accurate way to predict a geothermal system's historical mass extraction data objectively. This method relies on fitting an arbitrarily high order polynomial for each wells mass production. The approach calculates the weights of the terms in the polynomial such that the fit to the data is minimized. The uncertainty of model predictions can be quantified through Monte Carlo simulation uncertainty propagation. The second method uses Gaussian Process Regression (GPR) to solve this sparse data problem. GPR is a Bayesian approach that assumes a time correlation between dense data points based on a kernel function while respecting the sparse data for each well. This approach was modified so that the sum of wells also respected the dense time history data measured at the separator. Both approaches were tested on synthetic data and data from an operational geothermal field. The results of these methods are compared in this paper, but both show merit in providing solutions to this problem.

1. INTRODUCTION

Understanding where the mass extracted from a geothermal field is coming from is essential for understanding how to manage that geothermal field in the future. Over-extraction may reduce the life span of the geothermal field (Rybach, 2007). Often, regular measurements are taken once the fluid has been grouped in the surface network. A simplified diagram of an example surface network is shown in Figure 1. The mass or steam may be measured at the separator or turbine daily, but the contribution from the individual wells is likely only measured periodically through flow testing. When conducting reservoir modelling, one modelling stage is calibrating the model to production data. This involves matching observed transient datasets (such as pressure and enthalpy changes in the reservoir) by extracting mass from the reservoir and reinjecting it into model blocks, replicating what has historically occurred. Therefore, understanding the history of extraction and reinjection of fluid in the geothermal reservoir is fundamentally important for the reliability of reservoir models.

Geothermal is a data-poor industry, particularly when it comes to understanding the subsurface. Limited data sets arise from the sparse nature of wells and the expense associated with collecting data. This paper will not discuss how well data is collected, for that we refer the reader to Zarrouk and McLean, 2019. Although data analytics has been applied to various geothermal problems (Abrasaldo et al., 2024), it has not yet been formally applied to the problem presented in this paper. A significant focus in modern statistical tools, is on generative AI and machine learning, which rely on vast amounts of data to give effective results. Trying to predict the dense history of a geothermal well's production from limited data has traditionally been a manual iterative task. This is where an expert analyses the data of individual wells and decides on time-varying proportions of the total production. By applying these proportions, the total mass of a group of wells is divided, noting that the total mass at a separator or generator is measured more frequently (often daily). This is an iterative process, usually done in a large spreadsheet, where the expert tries to minimize the gap between the dense predictions of each well historical mass take and the measured data points. A flow chart of this process is shown in Figure 2. This process often takes weeks of analysis due to the inherent complexities of the history of the surface network. For accurate predictions, one first needs to understand how wells are grouped and a schedule of when each well is on and off. A problem arises when the grouping of wells changes over time, for example when wells are connected to multiple power units. This manual process by an expert is akin to a heuristic optimization problem where proportions are iteratively changed to minimize an objective function (the gap between the guess and the data).

In this paper we present two approaches to solving this problem mathematically. The first approach is a traditional optimization approach termed Tikhonov Regularized Linear Least Squares Optimization (TRLLSO), where we treat the solution of the contribution of a well as an arbitrarily high-order polynomial and find the optimal set of coefficients of the polynomial that minimizes the gap between the polynomial and the data. The second approach in this paper is applying Gaussian Process Regression (GPR), an approach derived in the Bayesian statistical framework. Here, we convert a time-varying prior distribution to a posterior distribution by specifying the data that

the posterior distribution must respect. Within this approach, we can specify how certain we are about the data, and the results provide uncertainty intervals where the data is sparse.

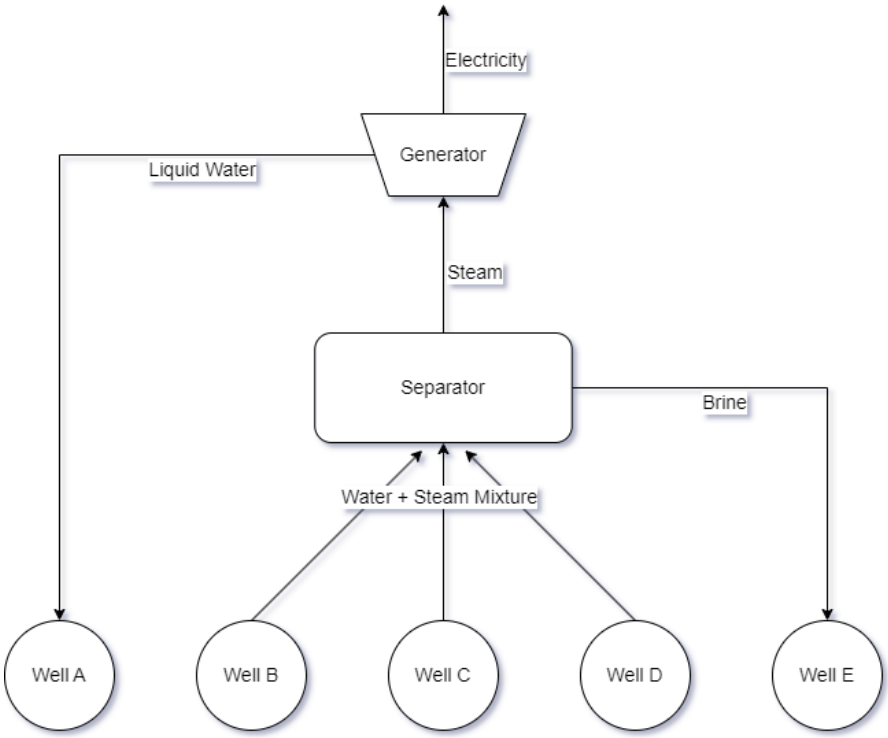

Figure 1: Simplified diagram of a surface network of a geothermal field.

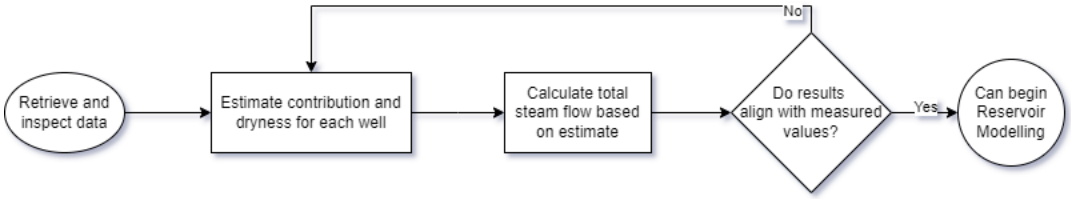

Figure 2: Iterative manual process for estimating individual wells contribution to the total production of a geothermal field.

2. METHODOLOGY

In this section we will discuss the methodology of the two approaches presented in this paper. Full details are omitted, but we refer the reader to (Clark, 2024) and (Terekhin, 2024) for a more complete description of the mathematical models of TRLSSO and GPR respectively.

2.1 Tikhonov Regularized Linear Least Squares Optimization

The model presented in this paper is the culmination of adding complexity to step towards more realistic well behavior and surface configurations. We refer the reader to (Clark, 2024) for a more detailed description of this progression. First, we define the mass extraction of an individual well as an arbitrarily high order polynomial,

$$W_i(t) = \sum_{j=1}^{k+1} w_{i,j}\, t^{j-1}, \tag{1}$$

where, $W_i(t)$ is the mass extracted from well, i, in time, $w_{i,j}$ is the weight of the term j which will be optimized and k is the order polynomial being considered. This is formulated as a loss function with Tikhonov regularization,

$$\mathcal{L}_{Tik} = \sum_{t=1}^{T}\left(M_t - \sum_{i=1}^{n} z_{i,t} W_i(t)\right) + \lambda_1 \sum_{i=1}^{n}\sum_{t=1}^{T} d_{i,t}\left(m_{i,t} - W_i(t)\right)^2 + \lambda_2 \sum_{i=1}^{n}\sum_{j=1}^{T}\left(\frac{dW_i(t)}{dt}|_{t=j}\right)^2 + \lambda_3 \sum_{i=1}^{n}\sum_{j=1}^{T}\left(\frac{d^2W_i(t)}{dt^2}|_{t=j}\right)^2,$$

(2)

where $\mathcal{L}_{Tik}$ is the loss function using Tikhonov regularization, M_t is the total mass extracted from a group of wells at time, t, $z_{i,t}$ is a binary variable determining if a well is on or off at time, t, $d_{i,t}$ is a binary variable representing if a measurement was taken at well i at time t, $m_{i,t}$ is the value of that measurement if taken and $\lambda_{1,2,3}$ are regularization parameters that determine the weighting of the individual well measurements, the first derivative of the estimated well flow and second derivative of the estimated well flow respectively. Formulating as a Loss Function Optimization problem we get,

$$\widehat{w} = arg\ min\big(\mathcal{L}_{Tik}(w_1, w_2 \dots w_n)\big) \tag{3}$$

Where $\widehat{w}$ is the optimal set of weights for each polynomial for each well. This problem can be solved analytically (with symbolic packages – SymPy was used in this case) but to optimally pick regularization parameters we need to conduct a grid or metaheuristic search in the λ_1, λ_2 and λ_3 space. This method as it stands is too computationally inefficient to do so. As such, this problem is reformulated as an inverse problem with the general form,

$$y = A_1 A_2 x + e \tag{4}$$

where, y is a vector that contains all the observations, both for grouped data and data for individual wells, A_1contains information about when wells were on or off and when wells were measured, A_2 is a block diagonal matrix where each block contains the forward polynomial model such that $q = A_2 x$ is the solution we are looking for, the mass flow of individual wells at all times. Additional equations are added in this matrix structure to include the terms related to Tikhonov regularization, we refer the reader to (Clark, 2024) for more detail.

To conduct parameter tuning, a Particle Swarm Optimization (PSO) metaheuristic is used. This is an optimization algorithm inspired by swarming animals such as insects and fish (Wang et al., 2018). This metaheuristic will give good (but not necessarily optimal) results in a reasonable amount of time and can be easily applied to a black-box function. A fixed value of k = 5 was used for testing. The regularization parameters that gave the lowest RMSE for the simulated dataset was $\hat{\lambda}_1 = 160.13$, $\hat{\lambda}_2 = 5.52$ and $\hat{\lambda}_3 = 0.0$. With the optimal value of $\hat{\lambda}_3$ being zero, this indicates that the second derivative term of the Tikhonov regularization is unnecessary. In the results section we will apply this formulation to both synthetic data sets and data from a real geothermal system.

2.2 Gaussian Process Regression

Gaussian Process Regression (GPR) is a machine learning method that employs Gaussian processes to model data. A GPR model consists of Gaussian processes, which are defined as a finite collection of joint Gaussian distributions characterized by random variables (Rasmussen, 2004). A Gaussian distribution in one-dimensional space is given by:

$$X \sim \mathcal{N}(\mu, \sigma^2) \tag{5}$$

where X follows a normal distribution with mean μ and variance σ^2, where σ denotes the standard deviation. Since a Gaussian process requires multiple random variables, this distribution must be extended to a multidimensional case. A joint Gaussian distribution, also referred to as a multidimensional Gaussian distribution, is defined over multiple continuous random variables (Ghosh et al., 2018). The one-dimensional distribution above can thus be extended as:

$$X \sim \mathcal{N}(\mu, \Sigma) \tag{6}$$

where X is a normally distributed random vector, μ is the mean vector, and Σ is the covariance matrix. The joint Gaussian distribution is thus fully determined by these two parameters. A Gaussian process extends this concept further by defining the distribution through a mean function and a covariance function (Rasmussen, 2004). This is expressed as:

$$f \sim \mathrm{GP}(m, k) \tag{7}$$

where function f is distributed as a Gaussian process with mean function, m, and covariance function, k (Rasmussen, 2004). The function f represents a distribution over functions, allowing for the generation of sample functions that define the prior distribution. The prior distribution, in the absence of training data, provides insight into the potential behaviour of the function.

When data is incorporated, the prior distribution is updated to form the posterior distribution. For simplicity, let the posterior distribution be

$$f \mid \mathcal{D} \sim \mathrm{GP}(m_{\mathcal{D}}, k_{\mathcal{D}}) \tag{8}$$

where,

$$m_{\mathcal{D}}(x) = m(x) + \Sigma(X, x)^{\mathsf{T}} \Sigma^{-1}(f - m) \tag{9}$$

$$k_{\mathcal{D}}(x, x') = k(x, x') - \Sigma(X, x)^{\top}\Sigma^{-1}\Sigma(X, x') \tag{10}$$

are the posterior mean and posterior covariance matrix, respectively.

The effectiveness of the posterior distribution depends on the selection of an appropriate prior distribution. If the function's behaviour is unknown, a prior cannot be fitted, requiring GPR to be trained solely on the provided data. A well-defined Gaussian process model depends on appropriate choices of mean and covariance functions (Rasmussen, 2004). Some detail of the mathematical model such as the optimization of hyper-parameters by finding the minimum of the negative log marginal likelihood function have been omitted but one important detail is the choice of covariance function or kernel, k. Since a Gaussian process mathematically assumes that the variable is a continuum, rather that supply a co-variance matrix it uses a kernel which is a function which returns a co-variance matrix once the variables of interest are defined. In this case, that is the time discretization used. In this paper we use the Matern Kernal which can be defined as,

$$k(x_i, x_j) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{l}d(x_i, x_j)\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}}{l}d(x_i, x_j)\right) \tag{11}$$

where ν controls the smoothness of the learned function, l is the length scale, $d(x_i, x_j)$ is the Euclidean distance between x_i and x_j , K_{ν} is the modified Bessel function, and Γ(ν) is the gamma function. In this case the length scale, l, is a hyper parameter and has bounds of (0.1, 10) with the time scale being in months or days. An example of realizations drawn from an example prior Gaussian Process, and the posterior after conditioning on data is shown in Figure 3. The reason for applying GPR to the problem of inferring the dense data history of individual wells is there is the flexibility to enforce fitting the data from individual wells, but the method also quantifies the uncertainty when looking at time periods where no data was collected. While the mean estimate could be used for traditional reservoir modelling (O'Sullivan and O'Sullivan, 2016), the uncertainty bounds are useful when applying a Bayesian approach and looking at uncertainty quantification of geothermal reservoir models (Maclaren et al., 2020 and Dekkers et al., 2022). In this case, the realizations from the posterior could be used as uncertain inputs to an ensemble of reservoir models where we assume we are uncertain about the history of each geothermal wells production.

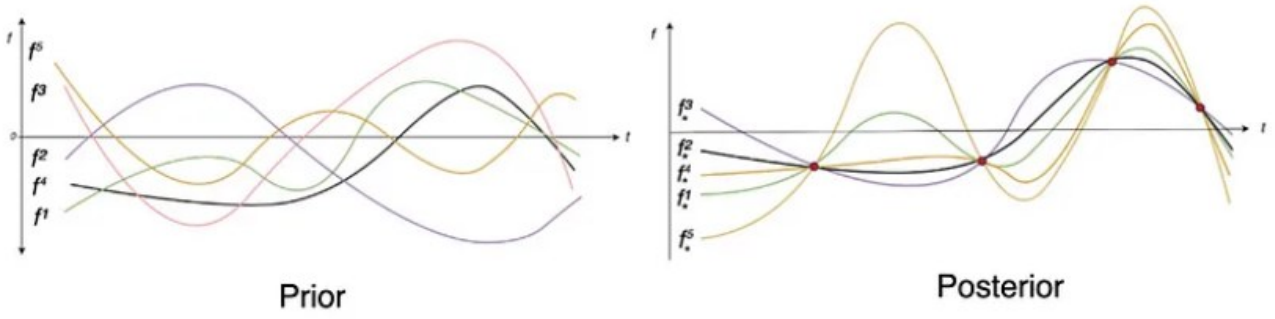

Figure 3: A simple demonstration of prior samples and posterior samples from their respective distributions from Gaussian Process Regression. Adapted from Hui, 2022.

3. RESULTS

In this section we will show the results of the TRLLSO and GPR methods. Both models were first tested and verified on simple synthetic data sets. We progressively relaxed the assumptions about the simplicity of the synthetic data and finally we tested each approach on a real data set where measurements were taken from an operational geothermal field in New Zealand. Both TRLLSO and GPR were supplied sparse individual well measurements but dense total or grouped measurements (i.e. the mass flow coming in that represents a group of wells such as at a separator or generator). For each model that time domain was discretized and it was assumed that there was a grouped measurement at every time period. It was also assumed that sparse measurements for individual wells were representative for the whole time period they occurred in.

Synthetic data comes from a synthetic reservoir model that is used for teaching purposes. The system is designed to represent a generic volcanic geothermal system in New Zealand (Renaud et al., 2022), and the well mass flows aim to replicate that. The original source of the synthetic data came from running a reservoir model with wells on deliverability. This meant that wells naturally declined in time and were affected by nearby production wells. Each method was supplied a subset of the data for individual well flows. On the figures, it indicates "true data" which is the full set of individual well data derived from the wells on deliverability. In this case the sum of the "true data" adds up perfectly to the grouped data. This represents an idealized condition, since in reality, the sum of the measurements at the well heads would not necessarily add to the separator total. There is also almost never a situation where all wells in a group are tested in the same time period. In Figure 7 the variable we are estimating and using measurements of is steam flow. This is functionally the same

as mass flow. In future investigations it would be valuable to include a combination steam flow and mass flow measurements to infer a wells production but that is not yet implemented.

For the real data set, the individual well data comes from Tracer Flow Tests (TFTs) which have been verified as sensible by a reservoir engineer at Contact Energy. Instead of "true data", we do not know what each well produces on a daily (or monthly basis). Instead, we plot an "expert's estimate". This is obtained by an "expert" manually solving the problem that we are solving with the proposed algorithms. The standard approach to solving this problem, is deciding on proportions that each well contributes to the total flow based on the individual measurements and divide the total flow up. This process is iterative as the proportions of wells will vary in time as individual well behavior varies.

3.1 Tikhonov Regularized Linear Least Squares Optimization

In Clark, 2024, a series of more simple problems were solved stepping towards the complexity shown in this section. Here we are fitting a 5th order polynomial to all cases using Tikhonov regularization where near optimal weights have been determined via a Particle Swarm Optimization meta-heuristic. In the simpler cases, we explored fitting lower order polynomials (such as constants and linear functions) and varied regularization terms but converged on this formulation as the results produced are reasonable for the set of test cases.

Figure 4 shows how TRLLSO performed on a synthetic data set. On the left (A) is the model estimates for the 5 individual wells. On the right (B) is a comparison of how the total modelled estimated mass flow matched the synthetic total mass flow. In the formulation of Tikhonov regularized Linear Least Squares Optimization, it explicitly includes whether a well is on or off and it can be seen that the model estimates respect this. The algorithm performed well given that it could see the * data points in (A) and the green line in (B).

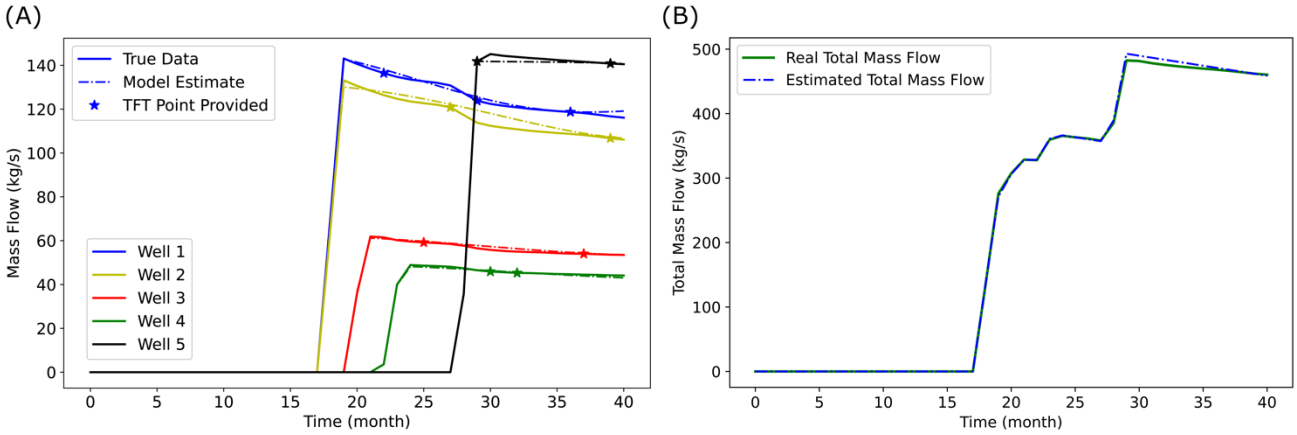

Figure 4: (A) Mass flows predicted by LLSO for simulated data, for k = 5 and optimal parameter values found using particle swarm optimization. (B) Total estimated mass flow from TRLLSO and synthetic total mass flows.

Next, we apply TRLLSO to the real data set, the results are shown in Figure 5. The algorithm provided relatively robust results however, as the complexity of the data increases, the ability for a 5th order polynomial to fit it is limited. For example, the model estimate is a poor fit to the second data point for Well 2 (yellow) in Figure 5. In this case the "experts estimate" is able to better represent this data point, but applying a trend that the well is declining doesn't reflect the fact that the TFT measurement increased from the first time Well 2 was measured to the second. Also, the experts estimate for Well 3 (red) increases (at approximately day 1000) after a period of time where it was shut off. No TFT point is provided over this period so TRLLSO has no information to indicate the flow from Well 3 should increase, but possibly the experts estimate was informed by additional anecdotal information such as the well being worked over during the shut down period, increasing its productivity. Inherently there will be trouble fitting all the available data, especially as real-world data sets often contain contradictions.

To compare to Gaussian Process Regression, and for the uncertainty quantification of reservoir models we were also interested in quantifying the uncertainty in the production data. For this model we applied Monte-Carlo Simulation where we assumed the TFT measurements had error associated with them. To simulate this, we applied Gaussian noise to the TFT measurements, with a mean of zero and a standard deviation of 5 kg/s. From here 2000 samples were taken and the optimal solutions found. The uncertainty bounds are shown in Figure 6. For the most part the 25th – 75th percentile of the LLSO estimates contain the experts estimate. The two areas where it does not contain the experts estimate are the same areas discussed above.

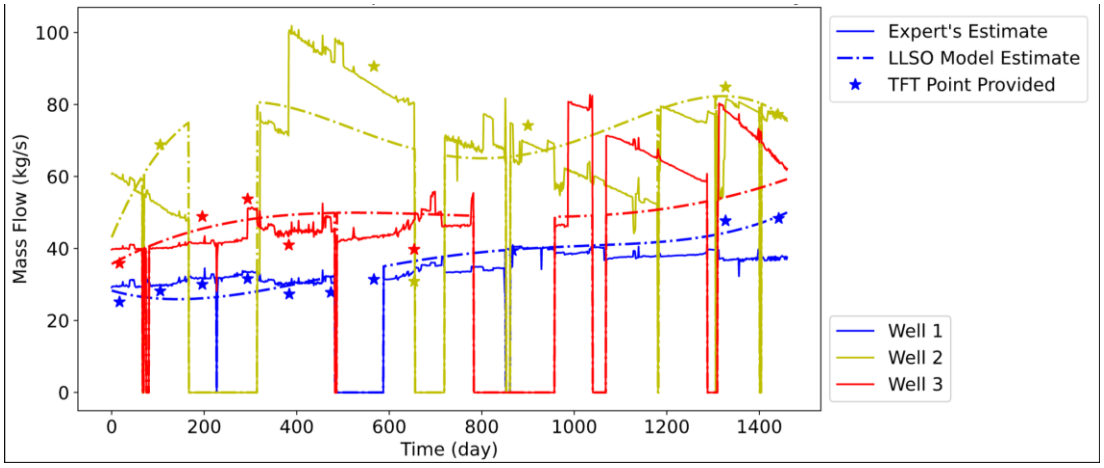

Figure 5: Mass flows predicted by LLSO for real data, for k = 5 with optimal regularization parameter values. Real data is for three wells feeding to the same separator at geothermal field in New Zealand. The experts guess is the current input to the reservoir model for this field and is obtained from manually fitting trends to each well. Individual well measurements were obtained through Tracer Flow Tests (TFT).

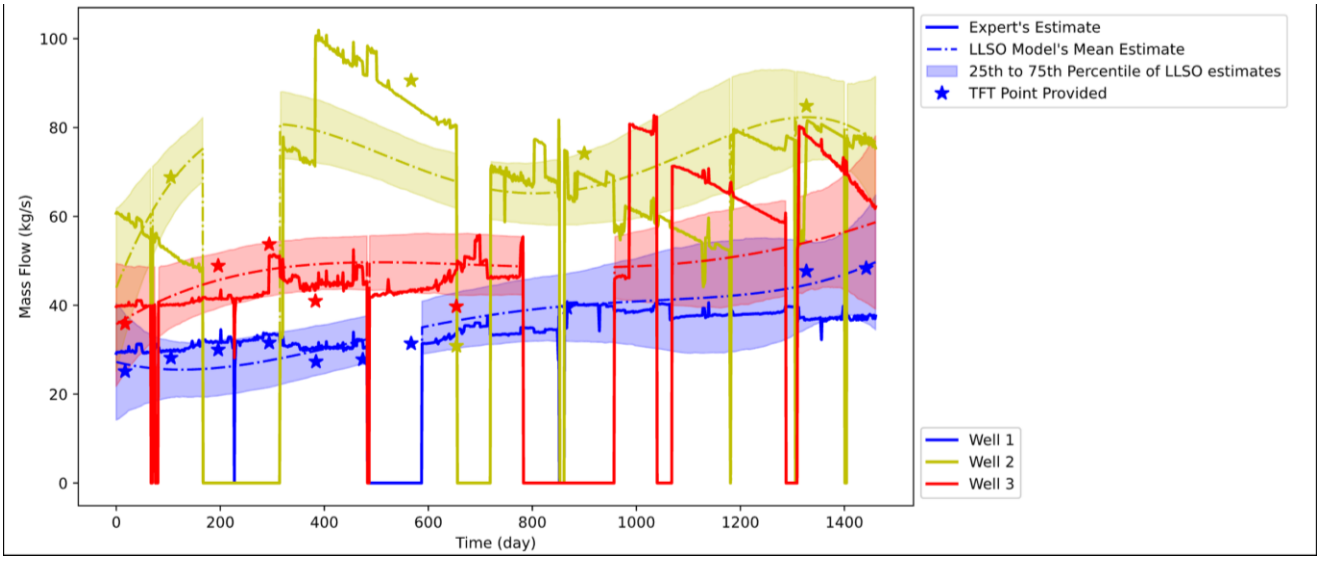

Figure 6: Monte Carlo simulation of model found in Figure 5. Gaussian noise with mean, 0, and standard deviation of 5 kg/s was added to TFT points. 25th to 75th percentiles of mass flows predicted by LLSO for real data.

3.2 Gaussian Process Regression

While aiming to find a solution to the same problem, Gaussian Process Regression has different underlying assumptions. Firstly, the sum of the wells explicitly equals the measurement of the total. For all realizations of the solution this condition is met. Also, the measurements of individual wells are also explicitly met by the solution. Noise (or error) in the measurements either of grouped data or well data can be incorporated. Doing so, will accept solutions that are within the given error tolerance. In all examples presented in this paper, we have assumed there is no error in the measurements. The second difference is that, unlike TRLLSO, in the formulation there is no formal way of describing a well as on or off. To handle the case where a well is not flowing, we assume that the well was measured to have zero flow and hence incorporate it like an individual well measurement in the same way we include Tracer Flow Tests. The third key difference is that there is no constraint currently in the model that says the mass flow must be explicitly positive, therefore sometimes the GPR solutions predict negative mass flows for wells that are meant to be producing. This is unphysical and will be improved in future iterations of this approach.

In Figure 7 we show the GPR solution for two synthetic data sets. The first (A), is a simple case where both wells are on for the full time domain. The uncertainty interval on the total steam flow (black line) is zero. This is because at each discretized time point, this data is known with zero error in measurement. Also as expected, for the second data point for Well 2, since we know the total, and we know the steam flow for Well 2, we must know the steam flow for Well 1. This is shown by the uncertainty interval for Well 1 collapsing to zero at this time. This indicates GPR is behaving as expected. The second synthetic data set is shown in (B) in Figure 7. In this case we

decompose the problem so we can make predictions for a well of interest where all other wells in the group are treated as one "lumped well" with no measurements. In this case,

$$q_{Total}(t) = q_{Well\ 1}(t) + q_{Lumped\ Well}(t), \tag{12}$$

where q is steam flow in Figure 7. For this synthetic case we supply GPR with measurements for the total steam flow at all time points and two measurements for Well 1. We also provide data points with zero steam flow when either Well 1 is off or all the wells that comprise other wells are off. Like the first example, when the steam flow for Well 1 is known, the problem is fully determined and the uncertainty interval for the lumped well goes to zero. Of note, after the last measurement for Well 1 (in month 30), GPR is extrapolating Well 1's behavior. In this case GPR performs poorly in comparison to the "true value" which comes from the deliverability model of the well. In light of this, it would be interesting to quantify the frequency of measurements required to obtain reasonable predictions.

Finally, the GPR approach was tested on the same real world data set used in Figure 5. Figure 8 shows that Gaussian Process Regression does adequately match the provided data. Despite the lack of a non-negativity constraint, the GPR model only predicts positive mass flows for this example data set. The second flow measurement for Well 2 is better captured by GPR compared to TRLLSO and as a result more closely aligns with the "experts estimate". Both TRLLSO and GPR fail to capture the increased production predicted by the expert for Well 3 after day 950. Again, this is likely due to the experts estimate including anecdotal evidence that Well 3 is expected to have more production. Another key observation is that in the early period (day 0 – 500), the density of TFT points for Well 1 and 3 is relatively high, which results in small uncertainty intervals on GPR predictions. In contrast to this from day 750 – 1300, the measurements are comparatively sparse and we observe higher uncertainty intervals. This aligns with intuition as the longer it has been since a well has been measured, the less certain one would be about predicting the flow rate of that well. This is a good indication that GPR may be a viable approach for this problem, albeit requiring some additional work to adequately constrain it.

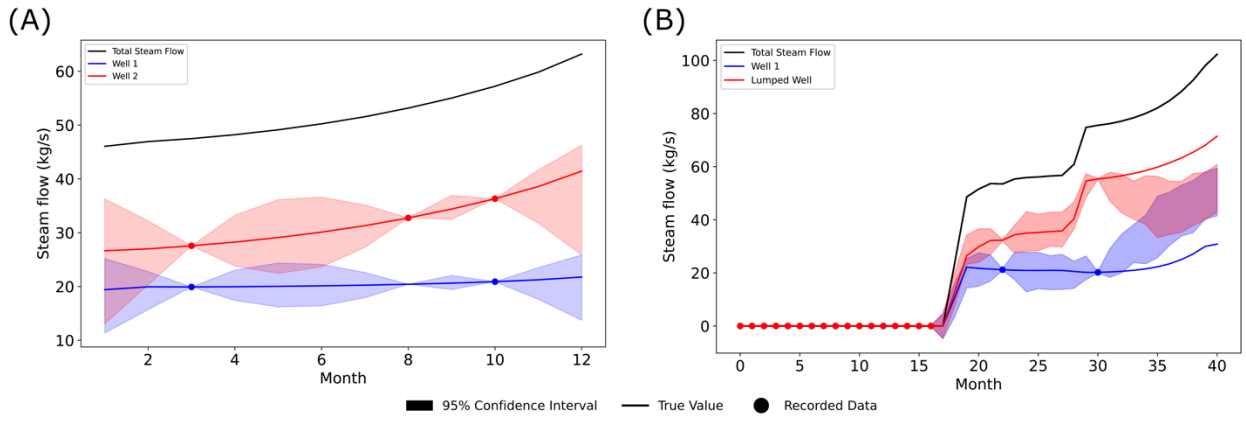

Figure 7: In each case synthetic data was used such that the model can see the "Recorded data" and the total steam flow (black line). (A) GPR model for two wells that are on for the duration of the 12-month period. (B) GPR model where a group of wells is treated like a lumped well and measurements for Well 1 are provided.

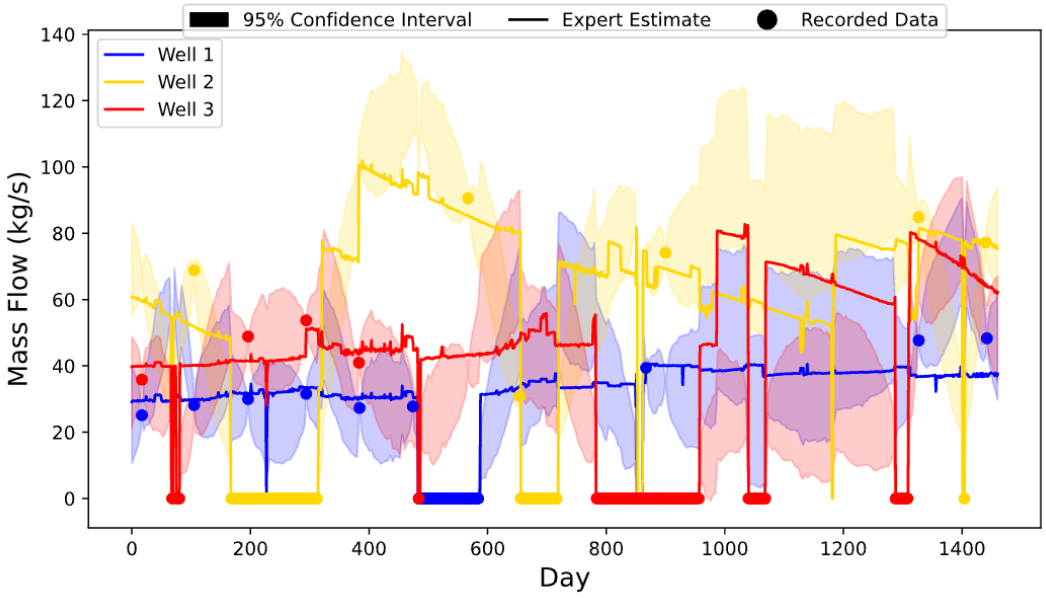

Figure 8: GPR model applied to real world data set. The data set and the experts estimate are the same as in Figure 5.

4. CONCLUSIONS

This paper presents a problem that is ubiquitous in the geothermal industry. Understanding the dynamics of a geothermal reservoir and how to manage the health of a reservoir is underpinned by first understanding where and when fluid was extracted historically. This information is required on a well-by-well basis for reservoir modelling but is only sparsely measured at the well. We have presented two approaches in this paper for inferring a dense well by well production history. The first approach was Tikhonov Regularized Linear Least Squares Optimisation (TRLLSO) and the second was Gaussian Process Regression (GPR). Both approaches were tested on synthetic and real data.

TRLLSO is solving an optimization problem by a fitting an arbitrarily high order polynomial to predict a time series mass flow for each well in the system. Tikhonov regularization was applied to reduce the over fitting of data while maintaining a physically realistic curve. Near optimal regularization weights were found using Particle Swarm Optimization and Monte Carlo Simulation was used to propagate uncertainty intervals onto the predictions of the model. On both synthetic data and real data for the problems presented here it performs well. Since it is solving a regression problem where it is finding the curves that are closest on average to all data points, on occasion it appears to omit a data point as the optimal solution favors matching other data. The predictions are restricted to the behavior of a polynomial and hence this approach may struggle if the data contains a large number of sharp changes, e.g. cycling wells. This may be a limitation of this method, and a larger, more complex data set may exceed the degrees of freedom we have when fitting a polynomial.

GPR is an algorithm set in the Bayesian statistical framework, where the goal is to find the posterior probability distribution for the mass flow of individual wells given that the solution passes through the provided data points. In the results presented here, it was assumed that the measurements had no error, however this framework is easily extendable to include uncertainty on measurements. On both the synthetic data and real data GPR was able to give sensible uncertainty intervals between data points which in contrast to TRLLSO would benefit from an increased data set size. GPR performed poorly at extrapolation, but perhaps in future this could be constrained by running wellbore simulations to constrain predictions of future well performance to realistic physical limits. The results of a wellbore simulations could be included as “data” with a suitable amount of uncertainty. The current clear shortfall of this approach is that we have not implemented a non-negativity constraint and hence in rare circumstances the method can produce negative mass flows which are not physical.

Overall, both TRLLSO and GPR show promise predicting a dense production history from sparse data. The data sets presented in this paper are limited in complexity so the next step will be moving towards incorporating the complexity we see in the real-world version of this problem. For example, a group of wells rarely stays constant through time, rather wells are switched between separators and units as the wells behavior changes over time. It would also be good if we could include grouped steam flow measurements to predict mass flow at a well. Often for a flash plant steam flow measurements are more readily available as they are important for understanding power plant efficiency. These methods could also be expanded to include predictions for injection mass flows as companies often do not have an accurate record of historical reinjection. We would also like to take the uncertainty bands produced by these methods and use them in the uncertainty quantification of geothermal reservoir models where we assume a wells production mass flow is uncertain, which given how sparsely it is typically measured is a reasonable step forwards.

ACKNOWLEDGEMENTS

The authors would like to thank Contact Energy Ltd. for supplying production data from a geothermal field in New Zealand.

REFERENCES

Abrasaldo, P. M. B., Zarrouk, S. J., Kempa-Liehr, A. W.,: A systematic review of data analytics applications in above-ground geothermal energy operations, Renewable and Sustainable Energy Reviews, Volume 189, Part B, (2024), 113998, https://doi.org/10.1016/j.rser.2023.113998.

Clark, D.: Historical Data Prediction for Geothermal Systems using Data-Driven Modelling. BHons thesis, Department of Engineering Science and Biomedical Engineering, University of Auckland, (2024). Available on request.

Dekkers, K., Gravatt, M., Maclaren, O., Nicholson, R., Nugraha, R., O'Sullivan, M.J., Popineau, J., Riffault, J. and O'Sullivan, J.P.: Resource assessment: estimating the potential of a geothermal reservoir. In: Proc. 47th Workshop on Geothermal Reservoir Engineering. Stanford University, Stanford, California, USA. (2022)

Ghosh, S., Das, N., Gonçalves, T., Quaresma, P., Kundu, M.: The journey of graph kernels through two decades, Computer Science Review, Volume 27, (2018), 88-111, https://doi.org/10.1016/j.cosrev.2017.11.002.

Hui, J.: Understanding Bayesian linear regression Gaussian process with Normal Distributions, (2022), https://jonathan-hui.medium.com/26bayesian-linear-regression-gaussian-process-with-normal-distribution-e686f7846ad1

Maclaren, O.J., Nicholson, R., Bjarkason, E.K., O'Sullivan, J.P. and O'Sullivan, M.J.: Incorporating posterior-informed approximation errors into a hierarchical framework to facilitate out-of-the-box MCMC sampling for geothermal inverse problems and uncertainty quantification. Water Resour. Res. 56 (1), (2020).

O'Sullivan, M.J. and O'Sullivan, J.P.: Reservoir modeling and simulation for geothermal resource characterization and evaluation. In: Geothermal Power Generation. Elsevier, pp. 165–199, (2016).

Rasmussen, C. E.: Gaussian Processes in Machine Learning, Springer Berlin Heidelberg, Berlin, Heidelberg, (2004), 63–71.

Renaud, T., Popineau, J., Riffault, J., O'Sullivan, J.P., Gravatt, M., Yeh, A., Croucher, A. E. and O'Sullivan, M.J.: Practical workflow for training in geothermal reservoir modelling. In: Proc. 43rd New Zealand Geothermal Workshop. Wellington, New Zealand, (2022).

Rybach, L.: Geothermal sustainabaility, Geo-Heat Centre Quarterly Bulletin 28, (2007).

Terekhin, M.: Inferring the detailed history of a geothermal field using sparse data and machine learning. BEHons thesis, Department of Engineering Science, University of Auckland, (2024). Available on request.

van de Schoot, R., Depaoli, S., King, R. et al., Bayesian statistics and modelling, Nature Reviews Methods Primers 1 (2021) 26pp. https://doi.org/10.1038/s43586-021-00017-2.

Wang, D., Tan L. and Liu, D.: Particle swarm optimization algorithm: an overview, Soft Computing, 22, (2018), 387–407. doi:10.1007/s00500-016-2474-6.

Zarrouk, S. J., Katie McLean, K.: Geothermal Well Test Analysis, Academic Press, (2019), 1-349, https://doi.org/10.1016/B978-0-12-814946-1.00015-3.