Comprehensive Output Estimation of Double Scattering Proton System with Analytical and Machine Learning Models

Jiahua Zhu
Rutgers-Cancer Institute of New Jersey, Rutgers-Robert Wood Johnson Medical School, New Brunswick, NJ

Taoran Cui
Rutgers-Cancer Institute of New Jersey, Rutgers-Robert Wood Johnson Medical School, New Brunswick, NJ

Yin Zhang
Rutgers-Cancer Institute of New Jersey, Rutgers-Robert Wood Johnson Medical School, New Brunswick, NJ

Yang Zhang
Rutgers-Cancer Institute of New Jersey, Rutgers-Robert Wood Johnson Medical School, New Brunswick, NJ

Chi Ma
Rutgers-Cancer Institute of New Jersey, Rutgers-Robert Wood Johnson Medical School, New Brunswick, NJ

Bo Liu
Rutgers-Cancer Institute of New Jersey, Rutgers-Robert Wood Johnson Medical School, New Brunswick, NJ

Ke Nie
Rutgers-Cancer Institute of New Jersey, Rutgers-Robert Wood Johnson Medical School, New Brunswick, NJ

Ning Yue
Rutgers-Cancer Institute of New Jersey, Rutgers-Robert Wood Johnson Medical School, New Brunswick, NJ

Xiao Wang (xw240@cinj.rutgers.edu)
Rutgers-Cancer Institute of New Jersey, Rutgers-Robert Wood Johnson Medical School, New Brunswick, NJ

Research Article

Keywords: treatment planning system (TPS), machine learning,

Posted Date: April 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-456327/v1

Version of Record: A version of this preprint was published at Frontiers in Oncology on January 31st, 2022. See the published version at https://doi.org/10.3389/fonc.2021.756503.

Comprehensive Output Estimation of Double Scattering Proton System with Analytical and Machine Learning Models

Jiahua Zhu[1], Taoran Cui[1], Yin Zhang[1], Yang Zhang[1], Chi Ma[1], Bo Liu[1], Ke Nie[1], Ning Yue[1], Xiao Wang[1*]

[1]Rutgers-Cancer Institute of New Jersey, Rutgers-Robert Wood Johnson Medical School, New Brunswick, NJ

*Corresponding Author:
Xiao Wang, Ph.D., DABR
Assistant Professor
Department of Radiation Oncology
Rutgers – Cancer Institute of New Jersey
Rutgers – Robert Wood Johnson Medical School
Rutgers, The State University of New Jersey
195 Little Albany Street, New Brunswick, NJ 08901
Phone: 732-354-2926
Email: xw240@cinj.rutgers.edu

Abstract

The beam output of double scattering proton system is difficult to be accurately modeled by treatment planning system (TPS). This study aims to design an empirical method using the analytical and machine learning (ML) models to estimate proton output in a double scattering proton system. Three analytical models and three ML models using Gaussian process regression (GPR) were generated on a training dataset consisting of 1544 clinical measurements, and the accuracy of each model was validated against additional 241 clinical measurements as testing dataset. Two most robust models (polynomial model and the ML GPR model with exponential kernel) were selected, and these two independent models agreed with less than 2% deviation using the testing dataset. The minimum number of samples needed for either model to achieve sufficient accuracy (±3%) was determined by evaluating the mean average percentage error (MAPE) with increasing sample number, and the differences between the estimated outputs using the two models were also compared for 1000 proton beams with randomly generated range, and modulation for each option. These two models can be used as an independent output prediction tool for a double scattering proton beam, and a secondary output check tool for cross check between themselves.

Introduction

Proton therapy is rapidly becoming one of the primary cancer treatment modalities in the recent decade. The utilization of the Bragg peak plays a pivotal role in delivering the prescription dose to the target, while sparing the normal tissues by stopping the proton beam at the distal end of the target [1-5]. In order to cover the entire target with a desired dose, the pristine Bragg peak has to be modulated to the spread-out Bragg peak (SOBP) in terms of target size and depth [6-8]. Due to the complexity of proton beamline to form various SOBPs in a double scattering proton machine, it is hard to model the output accurately. Therefore, most proton centers with a double scattering beam system have to measure the output of patient specific proton beams in a water phantom to determine required machine output, mostly in terms of Monitor Unit (MU).

In order to obtain the output of a proton beam conveniently and verify the output measurement, Kooy et al. proposed a semi-empirical analytical method to estimate the output as a function of r=(R-M)/M, where R and M denote the beam range and modulation, respectively [9,10]. This formula implements a basic model as a function of r, and also corrects for the effective source position and the inverse square law. However, this model was sensitive to the definition of range and modulation [5]. A variation of 18% in output was observed at beam data with small modulation [11]. Therefore, Lin et al. proposed a parametrized linear quadratic model which defined r with a limited length of modulation [5,11]. With this correction, the relative errors of predicted outputs compared to measured values were less than 3%. Besides, the basic model of output in Kooy's method was also fitted by the fourth order Taylor polynomial multiplied by a range-related factor, which was close to unity [12]. A comparison was also conducted between Kooy's original method and the Taylor series approach. The result showed the predicted values in the Taylor series approach were closer to the measurements. Sahoo et al. comprehensively analyzed the determination of output from proton machine beamline [13]. Relative output factor, SOBP factor and range shifter factor were the primary factors to determine the output. The result also showed a good agreement to the measurement within 2% for 99% of those fields. However, this method required a large amount of measurements to verify the conversion from SOBP factor and range shifter factor to output, which was time-consuming and complicated. Machine learning (ML) models have also been used in output prediction [14]. Sun et al. compared the accuracy of output from machine learning and Kooy's method [5]. Up to 7.7% of relative error from Kooy's method was reduced to 3.17% by machine learning.

We propose three analytical models and three machine learning algorithms for output estimation. The analytical models include a polynomial fitting model, a linear fitting model, and a logarithm-polynomial fitting model, all with different equations for different options. The machine learning algorithms utilize Gaussian process regression (GPR) model with different kernels to test the accuracy of output estimations, with one single model for all options. The definition of R and M is consistent with the machine vendor's definition and the data is from our clinical beam measurement. The comparison between predicted and measured outputs was performed. In addition, the minimum number of beam data measurements needed for building a robust model is discussed, which can provide some insights for clinical implementation.

Methods

1. Introduction of the proton machine

Mevion (Mevion Medical System, Inc. Littleton, MA) S250 is a single room compact proton treatment system. The system implements a unique design that places the synchrocyclotron on the outer gantry to connect to the in-room gantry. This design shortens the beam transport distance and reduces the footprint to 191 m^2. The synchrocyclotron can generate proton beam with a maximum energy of 250 MeV, and the beam generated from synchrocyclotron will be directly transported into the beam shaping system and delivered to the target. The in-room gantry travels in a C shape with the gantry angle ranging from 355 degree to 185 degree. The nominal source to axis distance (SAD) is 200 cm. Combined with the six-degrees-of-freedom (6DoF) treatment couch driven by a robotic arm, various treatment sites can be treated using the system.

The proton machine utilizes a double scatter system to broaden the pencil beam and creates a uniform dose distribution with a beam shaping system. The beam shaping system includes primary and secondary scatters, one absorber and one range modulator, which spreads out the Bragg peak. There are two types of nozzles on the inner gantry, a large applicator (maximum 25 cm in diameter) and a small applicator (maximum 14 cm in diameter), respectively. A brass aperture mounted on the applicator shapes the proton beam to cover the target. A compensator mounted at the end of the applicator modulates the distal end of proton beam. There are 24 options with different beam ranges, beam modulations, and field sizes, as listed in Table 1. The first 12 options are large options to be used with the large applicator. The other 12 options are deep/small options to be used with the small applicator.

2. Output measurement

Due to the complexity of proton beamline in a double scattering system, Varian Eclipse treatment planning system (TPS) (Varian Medical System, Palo Alto, CA) doesn't provide MU directly for a proton beam. Instead, the output has to be determined manually for each clinical proton beam.

To determine the MU for a clinical proton beam, a verification plan was generated by copying the original clinical proton beam to a water phantom with the same proton energy fluence. Regardless of the setup in the original clinical plan, a consistent setup with SSD = 190 cm was used instead for the verification plan. The compensator in the original clinical plan was removed in the verification plan to reduce measurement uncertainty. A reference point was added to determine the dose at the mid-SOBP of the beam, and the measurements were conducted in water phantom at the mid-SOBP of the same proton beam with a Farmer chamber (IBA Dosimetry America Inc., Memphis, TN) at SSD = 190 cm. Attention was paid to the in-plane location of the reference point to ensure lateral charge particle equilibrium for accurate dose prediction. Sun et al. and Sahoo et al. demonstrated that field size effect is negligible with a field opening of at least 5 cm diameter [5,13]. If the verification point was blocked, it was shifted. The same setup was then applied in a water phantom for absolute dose measurement. The absolute point dose at the verification point of 100 MU was measured following IAEA TRS398 protocol [15]. Given the outputs were the same for both clinical beam and verification beam, the MU of the patient specific beam would be calculated by taking the ratio of the verification point dose from TPS to the measured output.

3. Analytical model-based output estimation

In order to validate and verify, and eventually replace the manual measurement, output models were built based on previous measurements. Analytical models using an empirical formula to convert from range/modulation to output were built for each option, based on 1785 proton clinical field measurements. 1544 clinical proton fields from 2015 to 2019 were categorized as training dataset and the rest (241 fields) as the testing dataset. Three analytical models were employed to estimate the output and compared to the measurement as reference.

1) Polynomial fitting model

The polynomial fitting model is an adaptation of Kooy's empirical formula. In Kooy's formula [10], output is a function of r = (R-M)/M. According to the vendor definition, R is defined as the depth at distal 90% of normalized percent depth dose and M is defined as the length between proximal 95% and the distal 90% of normalized percent depth dose. The basic model of Kooy's formula is expressed in Ferguson et al. [12] as

$$d/MU\big(r(R,M)\big) = \frac{CF\times\Psi_c\times D_c}{100/(1+a_0 r^{a1})} \times [s_0 + s_1(R-R_L)] \times \left(\frac{ESAD(r)-\Delta z_p}{ESAD(r)-\Delta z_p-\Delta z}\right)^2. \quad (1)$$

The first term of Eq.1 is the basic output prediction; the second term corrects the variation of output related to the virtual source position and the third term is inverse square related.

A polynomial equation of each option was fitted to replace the basic model in Eq. 2 [12]

$$d/MU\big(r(R,M)\big) = (p_0 + p_1 r + p_2 r^2 + p_3 r^3 + p_4 r^4) \times [s_2 + s_3(R-R_L)] \times \left(\frac{ESAD(r)-\Delta z_p}{ESAD(r)-\Delta z_p-\Delta z}\right)^2, \quad (2)$$

where s_2 and s_3 are the option specific fitting parameter.

In terms of the fitting data, Ferguson et al. listed the values of s_0 and s_2 in different options and those are very close to unity[12]. The s_1 and s_3 were found to be much less than s_0 and s_2, therefore the variation of second terms from unity could be negligible. The third term is only to correct the measurement position if the effective source is not located at the middle SOBP. Therefore, if the SSD, rather than SAD, is used to setup, Δz is always zero and the third term equals unity.

Therefore, the equation of output estimation can be approximated by a quadratic Taylor polynomial in Eq. 3.

$$D/_{MU} = a \times r^2 + b \times r + c, \quad (3)$$

where $r = (R-M)/M$, D denotes the dose and a, b and c are the fitting parameter.

2) Linear fitting model

The Linear fitting model estimates output as the function of logarithm of R/M (Eq.4). The rationale of choosing this model was to space out data points clustered in the low R/M region, as observed from the polynomial model. From polynomial fitting graph, it was observed that the output variations in the low R/M region (full modulation) were larger, with a lot more data points than the high R/M region (Fig. 1). This finding is consistent with what Sun et al. and Kim et al. reported [5,16].

$$\log\left({}^{d}/_{MU}\right) = k \times \log({}^{R}/_{M)} + b. \quad (4)$$

3) Logarithm-polynomial fitting algorithm

The logarithm-polynomial fitting algorithm is an independent model from the previous two models, since the variables in previous models are both related to R/M. To build a model with a different variable, while still keeping the model accurate, different approaches were made and the most accurate one was selected. In this model, the output is the function of logarithm(R)/logarithm(M) in Eq.5.

$${}^{d}/_{MU} = a' \times r'^{2} + b' \times r' + c', \quad (5)$$

where $r' = \log(R) / \log(M)$.

4. Machine learning-based output estimation

Different from analytical methods, machine learning (ML) methods do not need to model option by option. Instead, they use option number, beam range and modulation to predict the output. To test the efficacy and accuracy of ML modelling, three ML GPR models with different kernels, including exponential kernel, squared exponential kernel, and rational quadratic kernel were used for the output calculation [17].

GPR is a non-parametric Bayesian approach towards regression problems that can be utilized in exploration and exploitation scenarios [17,18]. It predicts the output data by incorporating prior knowledge and fit a function to the data. The kernels play very significant roles in the regression modeling and can map the features from the original values to the featuring spaces by involving the latent variables. After the training process, the obtained models were evaluated using the parameters from the testing sets. The exponential kernel was based on the assumption that the Euclidean distances between different data points were Laplacian distributed. The squared exponential kernel used another assumption that the Euclidean distances were Normal distributed. The Rational Quadratic kernel can be seen as a scale mixture of Squared Exponential kernels with different characteristic length-scales [17].

5. Robustness of models related to sampling numbers

The robustness of output models could be impacted by the number of data fed into the model. The evaluation of minimum number of data necessary for a robust model was conducted by comparing different model outputs with increasing number of inputs. Models were built with different sampling numbers, randomly selected from the original training dataset. The sampling numbers ranged from 10 to the number of training dataset. Each time a new model was generated, and the mean average percentage error (MAPE) was calculated to evaluate the differences between predicted outputs and the corresponding measurements. The comparisons were performed per option for polynomial models.

6. Difference between analytical and ML models

Analytical fitting models and ML models play independently in output estimation. Cross check of output is essential for verifying the accuracy and effectiveness of the two methods. 1000 random points within the range and modulation of each option were generated to estimate output by using two kinds of models, and the results were compared.

Results

The total clinical fields including training data and testing data were categorized into 24 options (Table 1). In this table, Options 4, 6,7,8,9,10,20,21, and 22 were mostly used in clinic with sample numbers larger than 80.

Table 1. The Statistics of all options

	max range (cm)	min range (cm)	max modulation (cm)	total field
Option 1	25.0	22.6	20.0	55
Option 2	22.5	20.9	20.0	40
Option 3	20.8	18.8	20.0	76
Option 4	18.7	16.8	18.7	99
Option 5	16.7	14.9	16.7	68
Option 6	14.8	13.2	14.8	81
Option 7	13.1	11.5	13.1	90
Option 8	11.4	10.0	11.4	98
Option 9	9.9	8.6	9.9	90
Option 10	8.5	7.3	8.5	86
Option 11	7.2	6.1	7.2	45
Option 12	6.0	5.0	6.0	21
Option 13	32.0	29.6	10.0	3
Option 14	29.5	27.1	10.0	12
Option 15	27.0	24.6	10.0	37
Option 16	24.5	22.1	10.0	49
Option 17	22.0	20.1	10.0	7
Option 18	20.0	17.8	20.0	19
Option 19	17.7	15.4	17.7	52
Option 20	15.3	13.3	15.3	105
Option 21	13.2	11.2	13.2	173
Option 22	11.1	9.1	11.1	123
Option 23	9.0	7.0	9.0	65
Option 24	6.9	5.0	6.9	50

1. Accuracy analysis of output estimation

A deviation of 3% was used as tolerance in clinical output estimation. Analytical fitting curve of Option 5 is presented as an example in Fig.1 to show the absolute error of modelling output relative to the measured value. The output is plotted as a function of R and M, with polynomial fit in Fig.1 (a), linear fit in Fig.1 (b), and logarithm-polynomial fit in Fig.1 (c). The red dashed line represents ±3% from predicted output. The blue scattered marks representing the real measurements are all within 3% of the predicted value in Option 5, indicating accurate prediction for three models. The coefficient of determination of each fitting curve is provided on Fig. 1.

The histograms of the relative deviation of all 24 options are categorized in Fig. 2. Compared to the other two analytical models, polynomial fitting model provided a better agreement with measurement data, with all deviations within 3%. In ML GPR models, exponential kernel showed a more accurate output estimation than the other two with less than 2% deviation from the measurement. In addition, the testing data were imported into analytical and ML GPR models to verify the effectiveness and accuracy of output estimation (Fig. 3). It was observed that polynomial fitting model still provided a good output estimation within 3% deviation, and all the ML models also exhibited deviation within 3%. To summarize, polynomial model and ML GPR model with exponential kernel showed best performance among all 6 models, with less than 3% deviation from all measurement data.

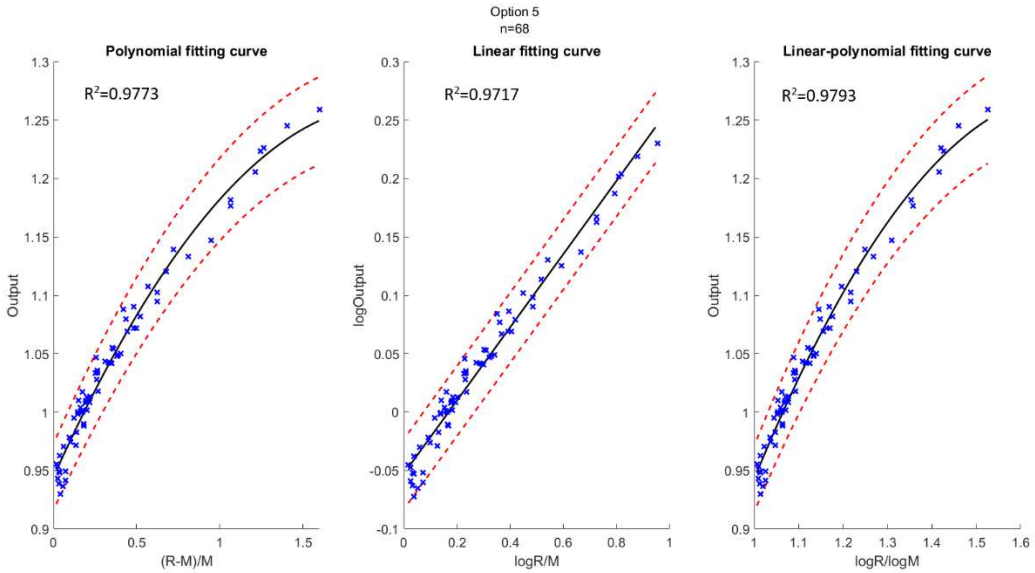


Fig. 1. Model based fitting curves for Option 5, including the polynomial fitting curve (a), the linear fitting curve (b), and the log-polynomial fitting curve (c). 3% confidence level in red dashed line. Number of data points n=68.

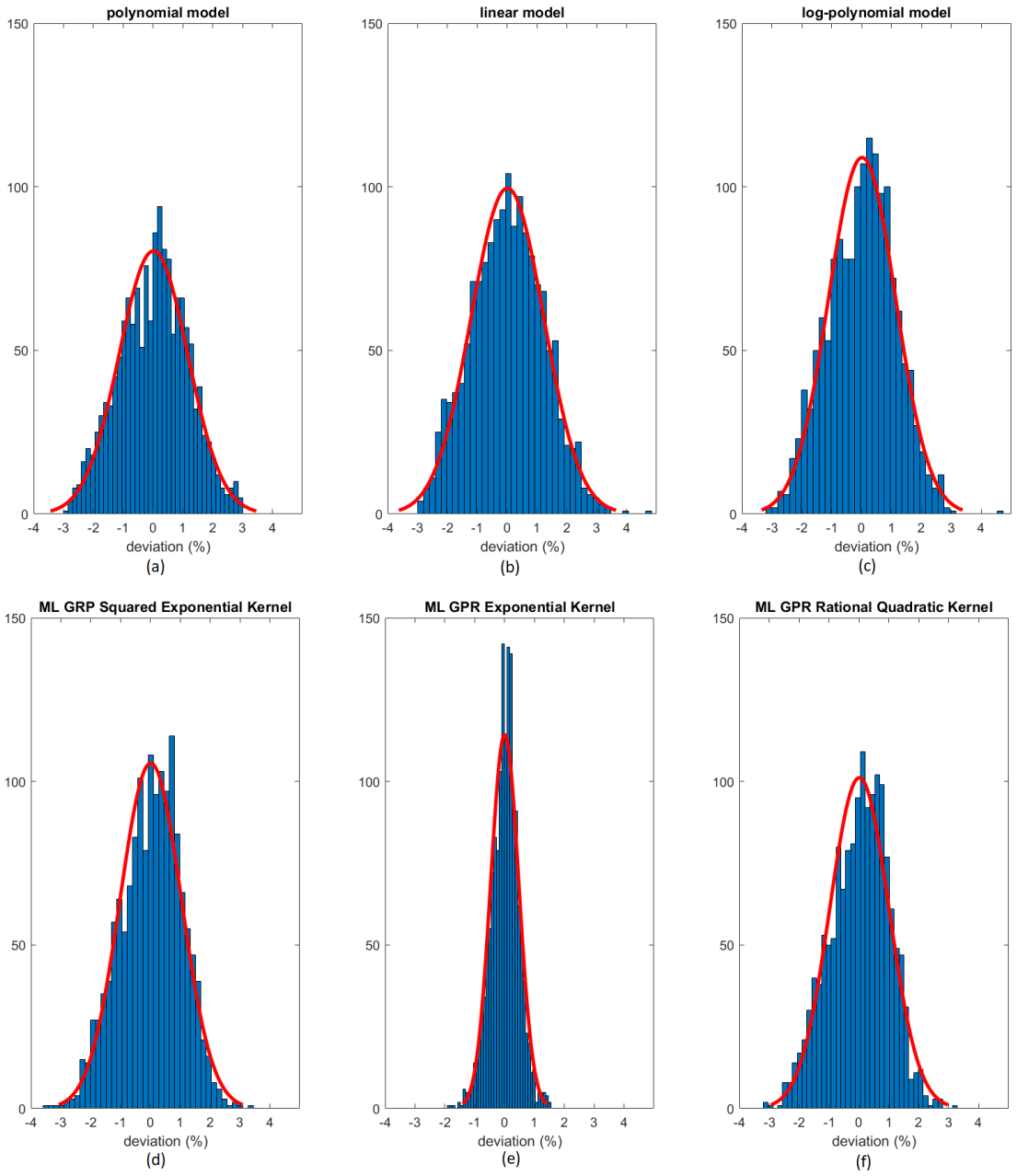


Fig. 2. Histograms of percent difference between analytical/ML GPR models and measurements using training data.

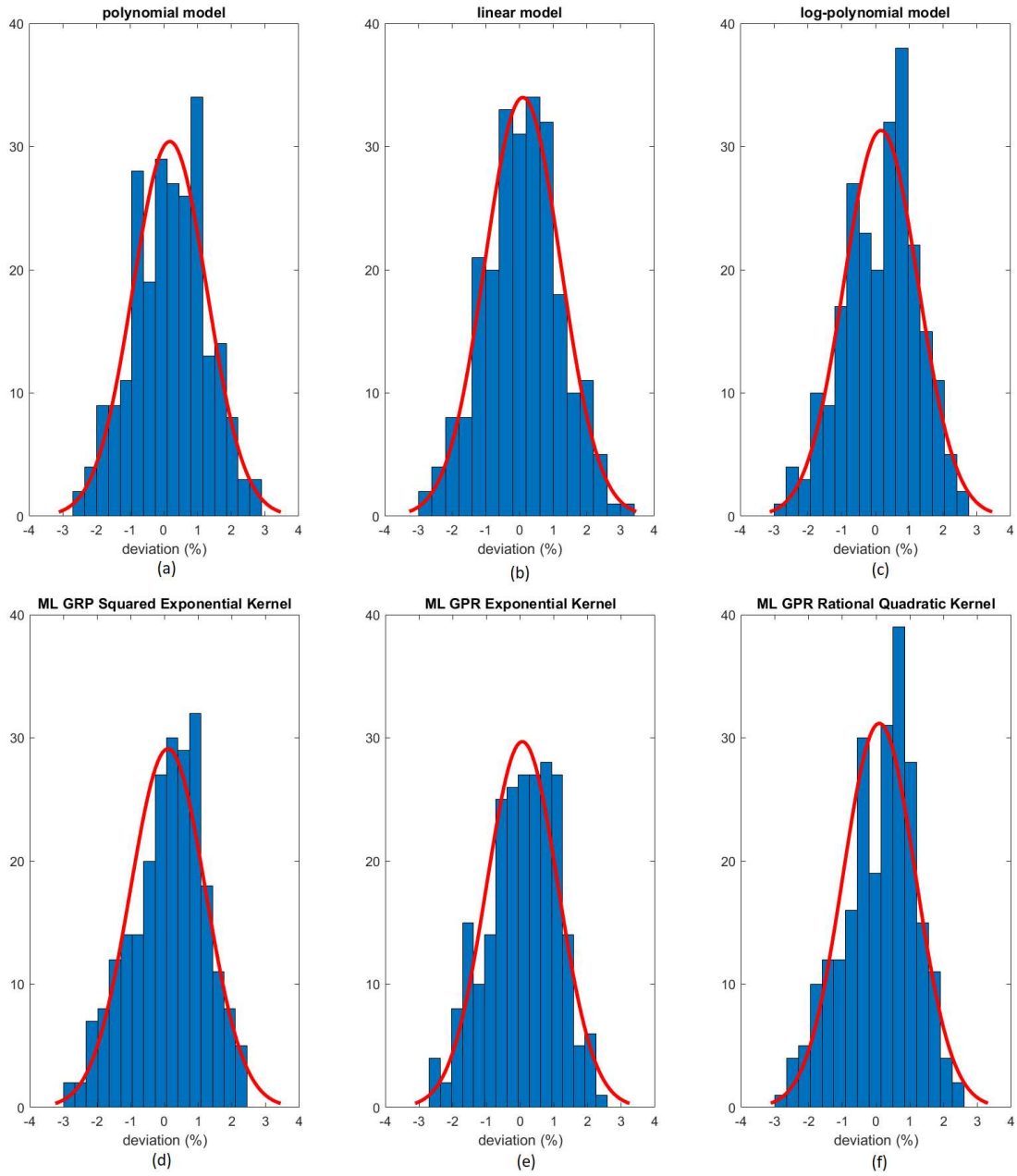
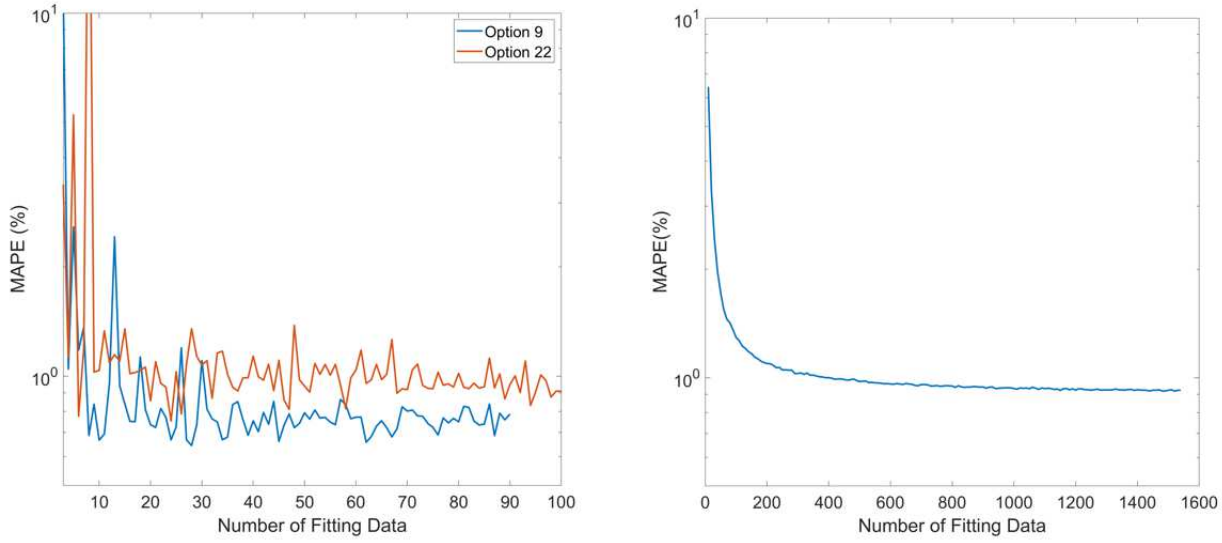


Fig. 3. Histograms of percent difference between analytical/ML models and measurements using testing data.

2. Minimum number of fields needed for polynomial model and ML GRP model with exponential kernel model

The trend of MAPE of models compared to measurements are shown in Fig. 4. Option 9 and Option 22 were chosen as the representatives of polynomial model because of higher sample numbers available, as shown in Fig.4 (a). The trend of MAPE for ML GPR model with exponential kernel is shown in Fig.4 (b). As observed in this figure, the relative error in Options 9 and 22 both converged to be around 1% or less once 20 data points were used

for building the polynomial model. For ML GPR model with exponential kernel, the convergence of MAPE was reached at around 400 data points, regardless of the option.



(a) (b)

Fig. 4. Mean average percentage error (MAPE) of polynomial model (a) and ML exponential kernel model (b) with the increase in fitting data.

3. Evaluation of difference between analytical models and ML models

Comparisons of output estimation between the polynomial fitting model and the ML GPR model with exponential kernel for Options 8 and 22 are shown in Fig. 5. MAPE for 1000 randomly generated points between the two models are shown with corresponding R and M. Measurement data are marked as pink scattered points overlaid on the figure. It can be seen from Fig. 5 that the two models agreed well in the regions where there was measurement data. Considerable differences in the outputs were observed beyond measurement region. More intuitive figures are shown in Fig. 5(b) where the general trend of outputs split in between polynomial and ML models with the decrease of modulation.

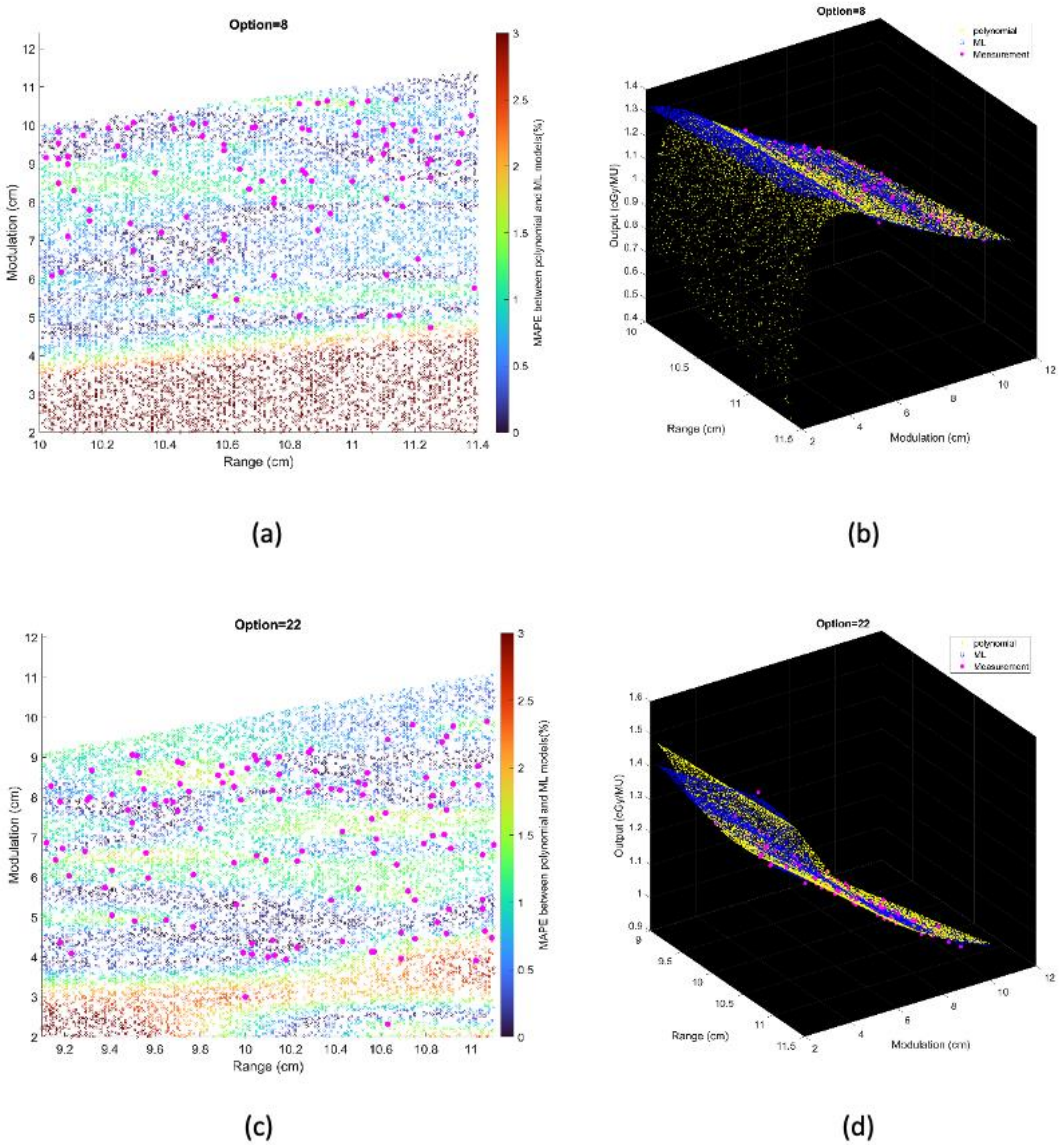


Fig. 5. The relative difference of output estimation with R and M. (a) and (c) show the differences of two models. The solid pink points are the measurement data, and the circles are MAPE between two models from the random data. (b) and (d) are the differences in 3D graphs that illustrate the trend of difference between two models.

Discussion

Currently the TPS cannot model the output for double scattering proton systems due to the complexity of proton beamline which varies with different options. Patient specific output estimation has to rely on manual measurement, which can be labor intensive and an error-prone process. To date there is no commercial software to verify manual measurement, so it is valuable to build models to second check manual measurement, and the ultimate goal of the study is to build an automatic output estimation and MU determination process. To achieve automation in output estimation, six models with different methods were built.

The output estimation derived from three analytical fitting models and three ML GRP models with different kernels were demonstrated and compared to the measurements. Polynomial fitting model and ML GPR model with exponential kernel with the best performance were chosen. In terms of the distribution in histogram, the polynomial fitting method provided the most accurate output estimation in analytical methods and ML GPR model with exponential kernel could provide more accurate output prediction than the other two ML GPR models. Also, the relative error between the estimated and the measured output for polynomial fitting model and ML GPR model with exponential kernel were always within ±3% in both training data and testing data. Therefore, it is proof that those two models could be adopted as the output estimation models. Also, since they are two independent models, it is suggested to they can be used as second check tools for clinical measurement, and also cross check tool for each other. The polynomial model is an expansion of Kooy's empirical formula using Taylor series. This approach is similar to the equation developed by Ferguson et al [9] but using a lower order of polynomials, thus simplifies the equation.

The number of data needed for establishing a robust polynomial model was estimated by the MAPE trend with increasing data points. Options 9 and 22 were shown as an example that the MAPE converged once the number of training data increased to 20. This gives a simple guidance on the number of data necessary to build an accurate polynomial model for an option. Among all options, some of them were rarely used, especially the deep options (option 13 with 3 beams, option 14 with 12 beams, option 15 with 37 beams, option 17 with 7 beams, and option 18 with 19 beams). This is because clinically we tend to plan the proton beam to penetrate through a shorter path if possible, leading to lower usage of deep ranged options. For those options with fewer data points, polynomial based output model would not be recommended. Instead, manual measurement would be required, until enough data points are accumulated.

For ML GPR model with exponential kernel, convergence of MAPE to 1% was observed after the input of 400 fields. This needs to be clarified that when building the ML models, range and modulation, as well as the option number were inputs to the model. Sun et al. also estimated the minimum number of field needed for ML cubist model [5]. In their study, the mean absolute error converged to 0.7% after 1200 data points. Their learning curves also showed a mean absolute error around 1% at 400 samples. Since the ML model doesn't discriminate different options, and some options may have fewer data samples than others, validation of accuracy of the model in all options is needed before clinic implementation.

Polynomial fitting model and ML method could be used as independent secondary check, and eventually the primary output estimation, replacing measurements. This requires the assessment of the agreement between two models. From Fig. 5, it is clear that the MAPE between two models was less than 2% if the data point lay within the region where there existed measurement data. Beyond this region (e.g. M=2-6 cm in Option 4), the two models showed obvious different trends with increasing differences, which indicated that the user must evaluate the accuracy of output prediction with extra measurements, otherwise the model cannot be used beyond the region with real measurements. It is suggested that the models should only be trusted to replace measurements with judgement that the beamline (R and M) falls within the region with enough measurement data.

Even though polynomial fitting model and ML GPR model with exponential kernel proved their feasibility for output estimation, it is still essential to pay attention to MU determination, as the MU is related not only to output, but also that its accuracy is related to the verification point dose. The accurate selection of verification point is pivotal in MU determination. It is recommended to perform a sanity check on the MU of clinical plan. A simple sanity check is to compare calculated MU to the prescription dose multiplied by the field weighting. The rationale of this sanity check is because the output is always close to 1. Future work includes automatic MU determination with Eclipse Scripting, to help get rid of the uncertainty of manual selection of verification point. Nevertheless, whether the output is measured or modelled, the MU must be verified prior to clinical treatment.

Conclusions

MU determination including output measurement is one of the most time-consuming and complicated work in patient QA for double-scattering proton machines. Comparing with the output measurement, analytical fitting models or ML models are more efficient to provide output estimation. Polynomial fitting method and ML GPR model with exponential kernel both show accurate estimation, and the accuracy meets the clinical requirement (within ±3%). These independent output estimations can serve as second check tools for measurements, and have potential to replace the measurement as part of the standard MU determination procedure.

Author Contributions

XW, YZ and KN led the conception and design of the study. JZ, XW, BL, YZ, TC and KN contributed to acquisition of data. JZ, XW, TC, YZ and CM contributed to analysis, data modelling and interpretation of data. JZ, XW, TC and NJY drafted and revised the article. All authors have approved the final article.

Competing Interests

The authors declare no competing interests.

Data Availability

The datasets generated or analyzed during the current study are available from the corresponding author on reasonable request.

Reference

1 Wilson, R. R. Radiological use of fast protons. *Radiology* **47**, 487-491 (1946).

2 Weber, D. C., Trofimov, A. V., Delaney, T. F. & Bortfeld, T. A treatment planning comparison of intensity modulated photon and proton therapy for paraspinal sarcomas. *International Journal of Radiation Oncology* Biology* Physics* **58**, 1596-1606 (2004).

3 Kim, T. H. *et al.* Efficacy and feasibility of proton beam radiotherapy using the simultaneous integrated boost technique for locally advanced pancreatic cancer. *Scientific reports* **10**, 1-10 (2020).

4 Rana, S., Simpson, H., Larson, G. & Zheng, Y. Dosimetric impact of number of treatment fields in uniform scanning proton therapy planning of lung cancer. *Journal of Medical Physics/Association of Medical Physicists of India* **39**, 212 (2014).
5 Sun, B. *et al.* A machine learning approach to the accurate prediction of monitor units for a compact proton machine. *Medical physics* **45**, 2243-2251 (2018).
6 Paganetti, H. *Proton therapy physics*. (CRC press, 2018).
7 Farr, J. *et al.* Clinical characterization of a proton beam continuous uniform scanning system with dose layer stacking. *Medical physics* **35**, 4945-4954 (2008).
8 Lansonneur, P. *et al.* First proton minibeam radiation therapy treatment plan evaluation. *Scientific reports* **10**, 1-8 (2020).
9 Kooy, H. M., Schaefer, M., Rosenthal, S. & Bortfeld, T. Monitor unit calculations for range-modulated spread-out Bragg peak fields. *Physics in Medicine & Biology* **48**, 2797 (2003).
10 Kooy, H. M. *et al.* The prediction of output factors for spread-out proton Bragg peak fields in clinical practice. **50**, 5847 (2005).
11 Lin, L., Shen, J., Ainsley, C. G., Solberg, T. D. & McDonough, J. E. Implementation of an improved dose-per-MU model for double-scattered proton beams to address interbeamline modulation width variability. *Journal of applied clinical medical physics* **15**, 297-306 (2014).
12 Ferguson, S., Ahmad, S. & Jin, H. Implementation of output prediction models for a passively double-scattered proton therapy system. *Medical physics* **43**, 6089-6097 (2016).
13 Sahoo, N. *et al.* A procedure for calculation of monitor units for passively scattered proton radiotherapy beams. *Medical physics* **35**, 5088-5097 (2008).
14 Kang, J., Schwartz, R., Flickinger, J. & Beriwal, S. Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective. *International Journal of Radiation Oncology* Biology* Physics* **93**, 1127-1135 (2015).
15 Radiology, I. A. E. A. W. H. O. P. A. H. O. E. S. o. T. & ESTRO., O. I. W. P. *Absorbed Dose Determination in External Beam Radiotherapy: An International Code of Practice for Dosimetry Based on Standard of Absorbed Dose to Water*. (International Atomic Energy Agency, 2000).
16 Kim, D. W. *et al.* Prediction of output factor, range, and spread-out Bragg peak for proton therapy. **36**, 145-152 (2011).
17 Rasmussen, C. E. in *Summer school on machine learning.* 63-71 (Springer).
18 David, B. Bayesian Reasoning and Machine Learning. *London: Cambridge* (2012).

Legends

Table 2. The Statistics of all options

Fig. 6. Model based fitting curves for Option 5, including the polynomial fitting curve (a), the linear fitting curve (b), and the log-polynomial fitting curve (c). 3% confidence level in red dashed line. Number of data points n=68.

Fig. 7. Histograms of percent difference between analytical/ML GPR models and measurements using training data.

Fig. 8. Histograms of percent difference between analytical/ML models and measurements using testing data.

Fig. 9. Mean average percentage error (MAPE) of polynomial model (a) and ML exponential kernel model (b) with the increase in fitting data.

Fig. 10. The relative difference of output estimation with R and M. (a) and (c) show the differences of two models. The solid pink points are the measurement data, and the circles are MAPE between two models from the random data. (b) and (d) are the differences in 3D graphs that illustrate the trend of difference between two models.

Figures

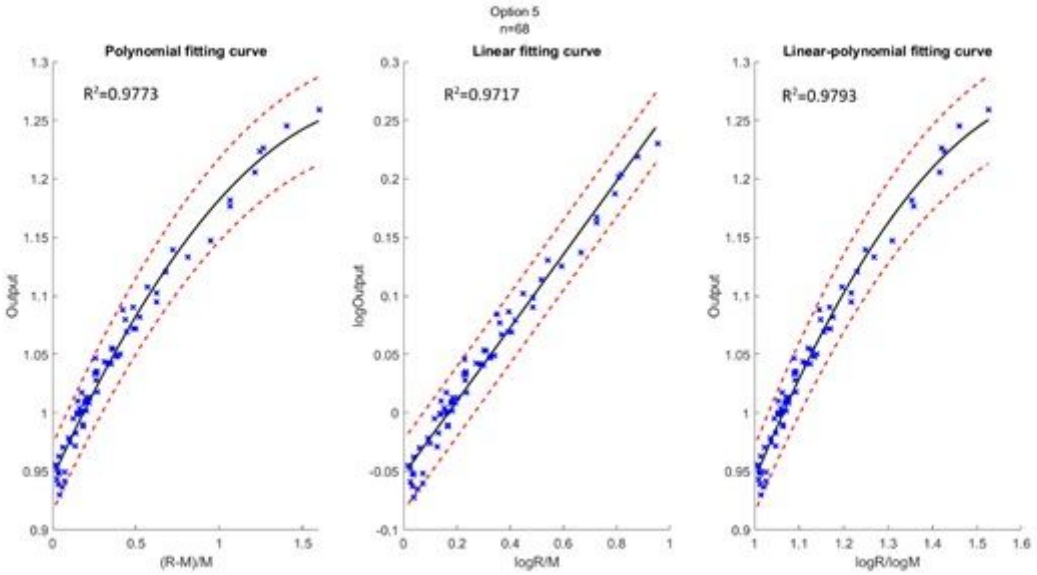


Figure 1

Model based fitting curves for Option 5, including the polynomial fitting curve (a), the linear fitting curve (b), and the log-polynomial fitting curve (c). 3% confidence level in red dashed line. Number of data points n=68.

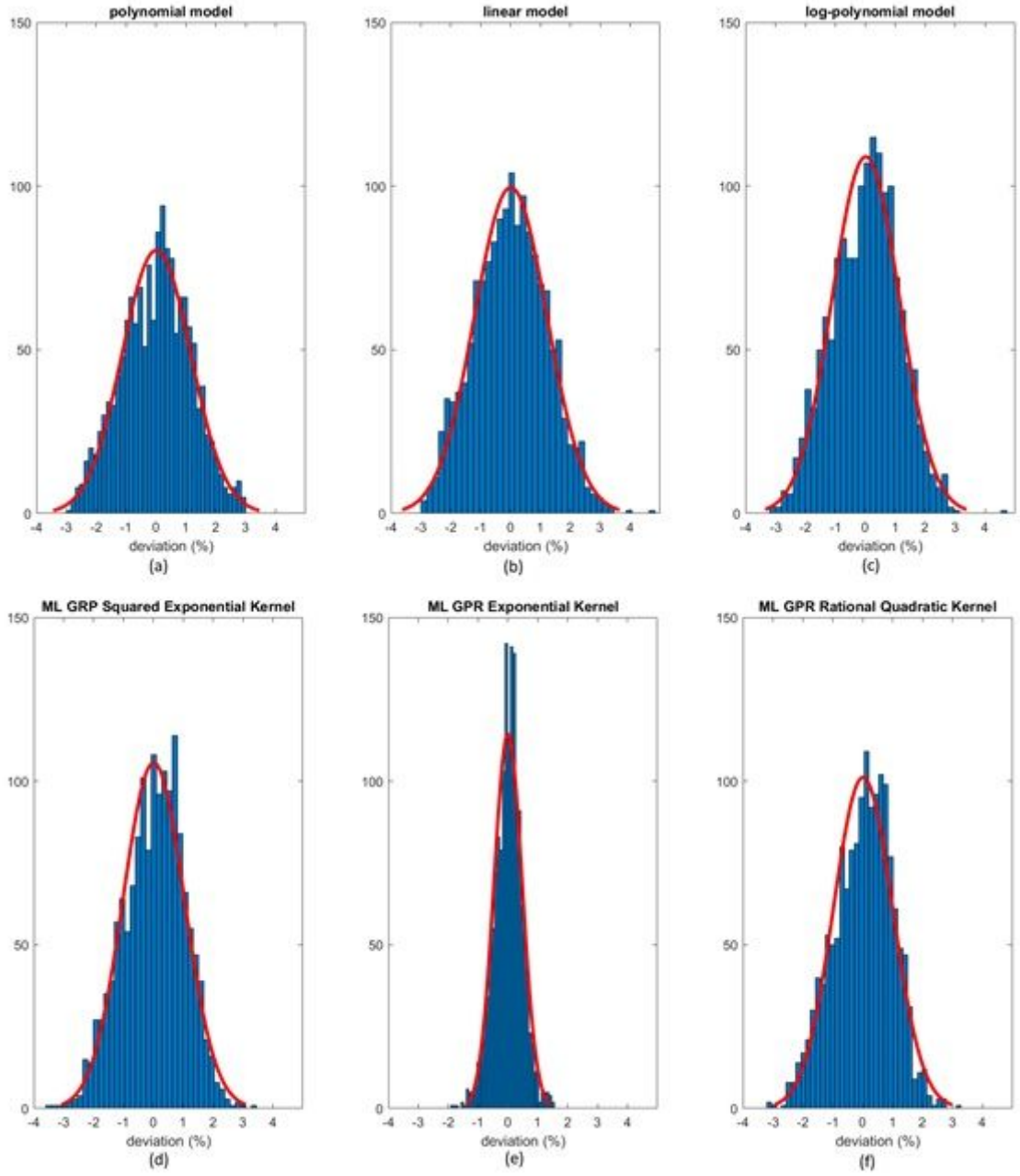


Figure 2

Histograms of percent difference between analytical/ML GPR models and measurements using training data.

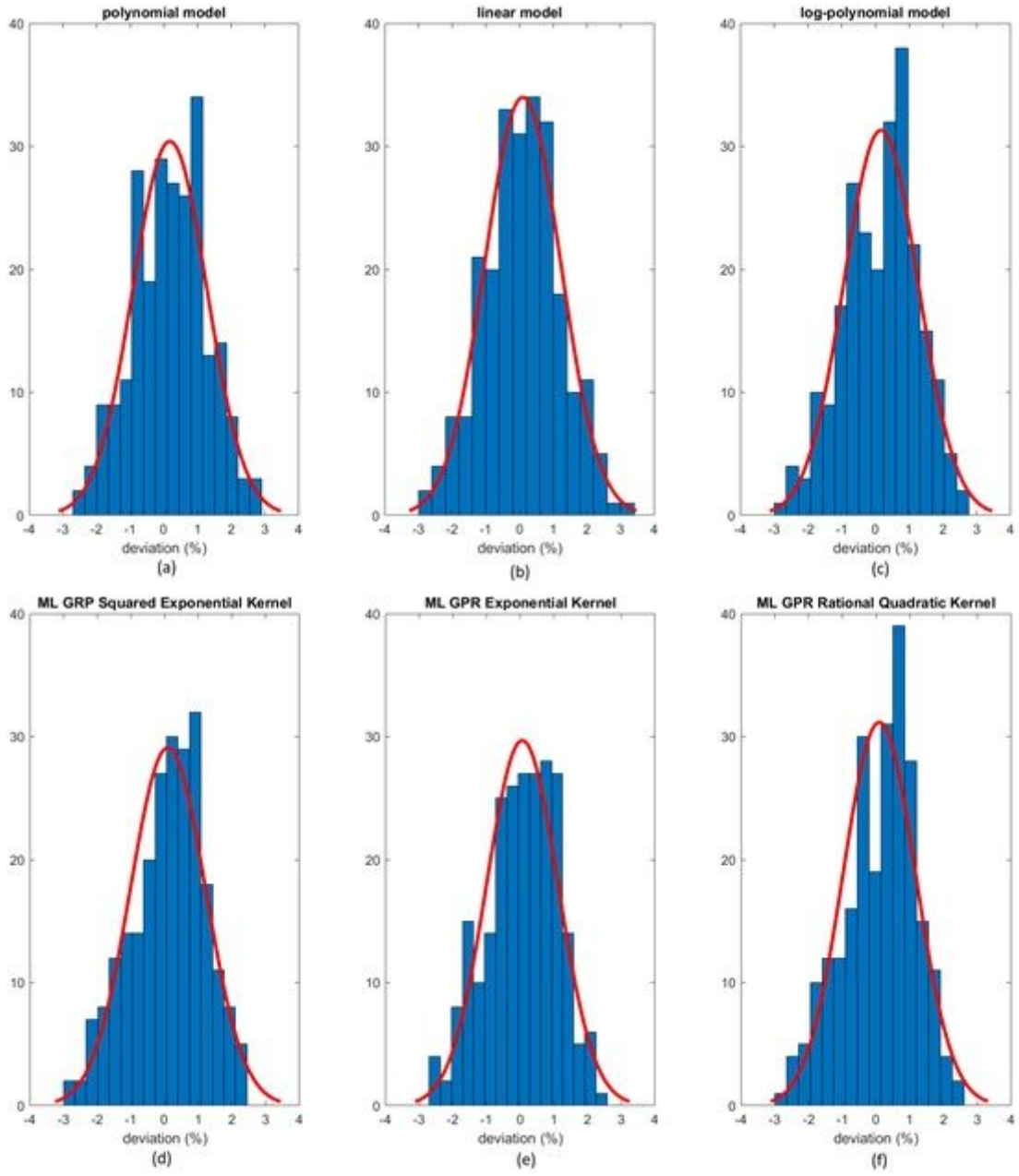


Figure 3

Histograms of percent difference between analytical/ML models and measurements using testing data. 2. Minimum number of fields needed for polynomial model and ML GRP model with exponential kernel model The trend of MAPE of models compared to measurements are shown in Fig. 4. Option 9 and Option 22 were chosen as the representatives of polynomial model because of higher sample numbers available, as shown in Fig.4 (a). The trend of MAPE for ML GPR model with exponential kernel is shown in Fig.4 (b). As observed in this figure, the relative error in Options 9 and 22 both converged to be around 1% or less once 20 data points were used for building the polynomial model. For ML GPR model with exponential kernel, the convergence of MAPE was reached at around 400 data points, regardless of the option.

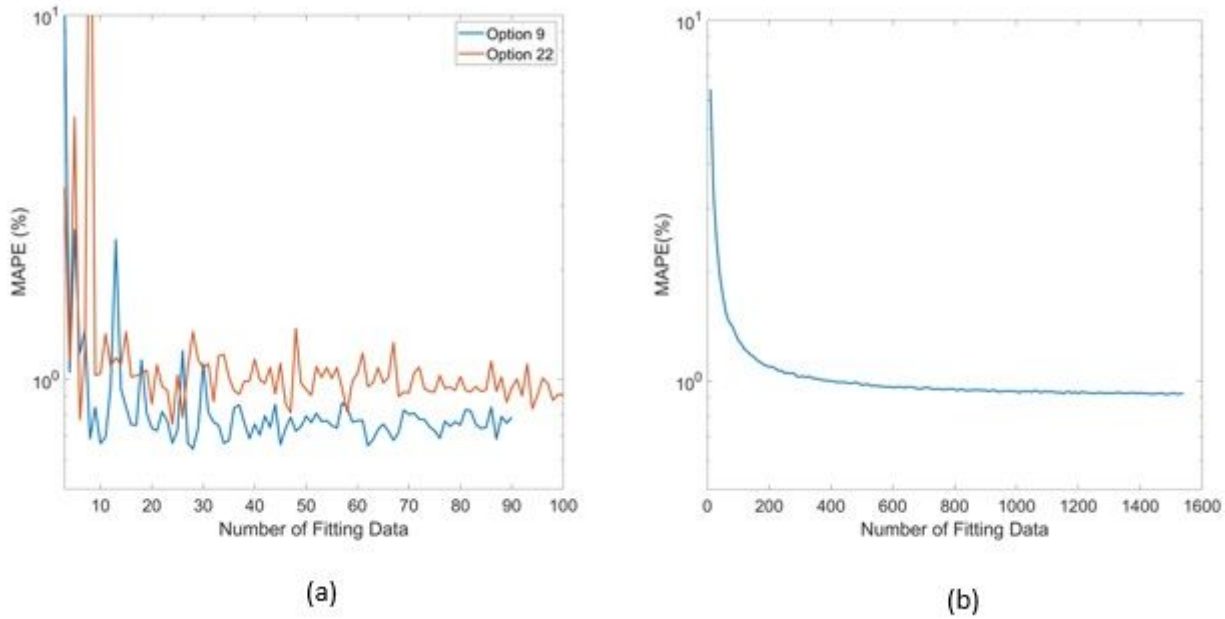


Figure 4

Mean average percentage error (MAPE) of polynomial model (a) and ML exponential kernel model (b) with the increase in fitting data. 3. Evaluation of difference between analytical models and ML models Comparisons of output estimation between the polynomial fitting model and the ML GPR model with exponential kernel for Options 8 and 22 are shown in Fig. 5. MAPE for 1000 randomly generated points between the two models are shown with corresponding R and M. Measurement data are marked as pink scattered points overlaid on the figure. It can be seen from Fig. 5 that the two models agreed well in the regions where there was measurement data. Considerable differences in the outputs were observed beyond measurement region. More intuitive figures are shown in Fig. 5(b) where the general trend of outputs split in between polynomial and ML models with the decrease of modulation.

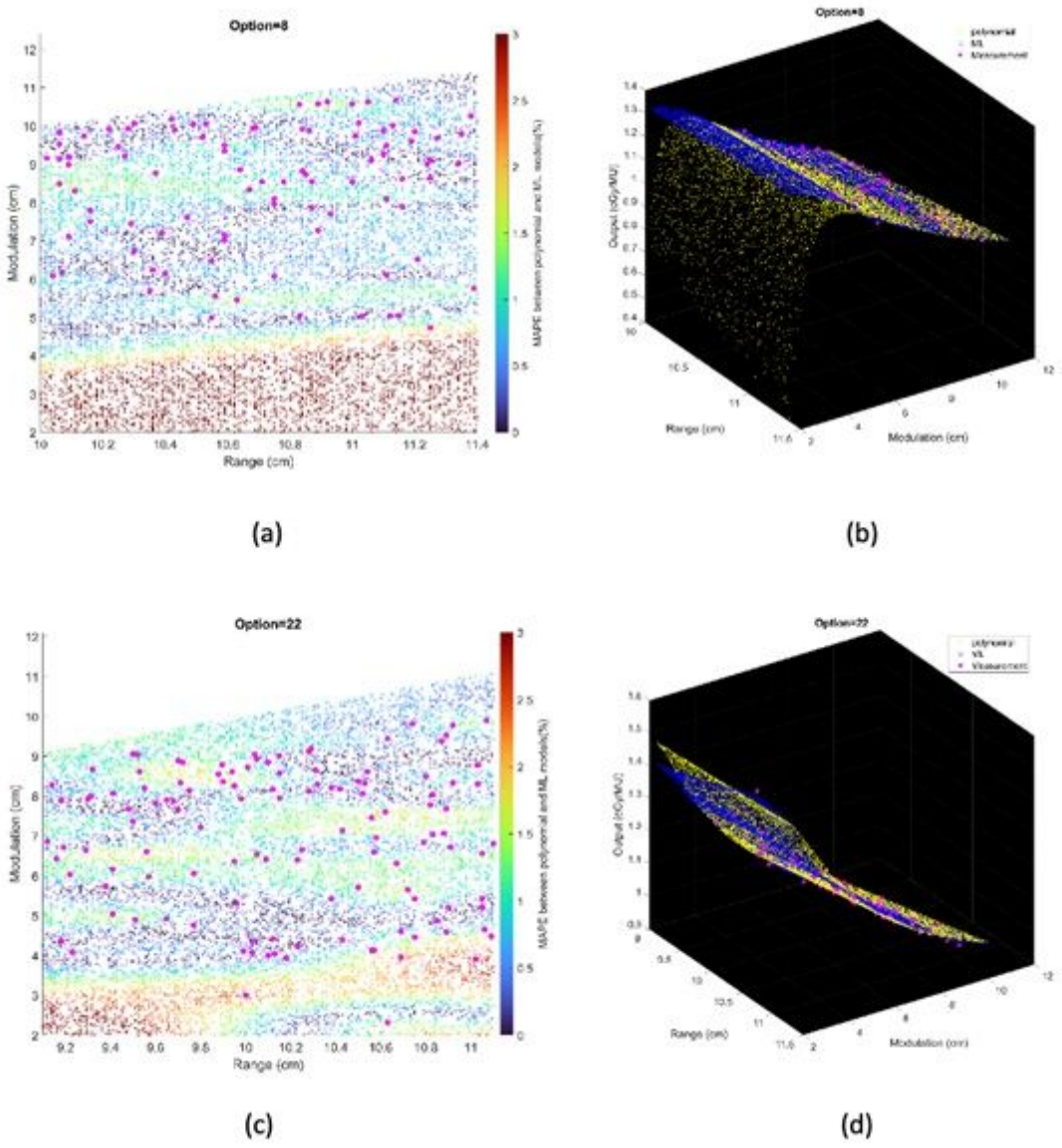


Figure 5

The relative difference of output estimation with R and M. (a) and (c) show the differences of two models. The solid pink points are the measurement data, and the circles are MAPE between two models from the random data. (b) and (d) are the differences in 3D graphs that illustrate the trend of difference between two models.