Forecasting models for the displacements and the piezometer levels in a concrete arch dam

By

Moneek Dilawari

March 2018

Department of Civil Engineering and Applied mechanics

McGill University, Montreal

Thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the the requirements of the degree of Master of Engineering

Abstract

Dam failures are termed as low-frequency & high loss events since the failure of a dam can lead to catastrophic results. Therefore, major dams in the world have been installed with dam monitoring systems which collect data regarding the behavior of the dam. This thesis aims to improve the forecasting models for the displacements & the piezometer levels in a dam. The Hydrostatic-Seasonal-Time (HST) is a statistical model which has been used by the dam owners to monitor and forecast the displacements of a dam to an external loading. However, the HST model is incapable of producing accurate results for non-linear relationships between variables, and cannot adequately account for changes due to excessively hot or cold seasons. In this thesis, three different variants of the HST model have been proposed to remove these defects from the analysis. A segmented HST model has been proposed which uses segmented regression to approximate the non-linear relationship between the variables. A non-parametric model has been proposed which replaces the regular seasonal component of the HST model with a non-parametric seasonal component computed using locally weighted linear regression. Finally, an HST_LT model is implemented which uses the lagged air temperatures around the dam as predictors to account for the fluctuations caused due to delayed thermal effects. The results suggest an improvement in the accuracy of prediction and a reduction in the scattering of the residual plots. Further, two models have been proposed to forecast the piezometer levels in the upstream and the downstream end of the dam. These models employ dam displacements as additional predictors along with the water level in the reservoir. The author has concluded that the relationship between the reservoir level and the piezometer level has hysteresis and that a polynomial expression is not sufficient to account for the effects of the reservoir level. The use of the displacements as additional variables improves the overall adjusted R^2 and reduces the standard error of the model. A segmented model

has been proposed to approximate the non-linear relationship between the displacements and the piezometer level which reduces the residual errors by almost 10% as compared to the regular model.

Resume

Les défaillances de barrages sont des événements à basse fréquence et à forte perte puisque la défaillance d'un barrage peut avoir des résultats catastrophiques. Les principaux barrages dans le monde ont donc été installés avec des systèmes de surveillance qui recueillent des données sur le comportement de chaque barrage. Cette thèse vise à améliorer les modèles de prévision pour les déplacements et les niveaux de piézomètre dans un barrage. Le modèle statistique hydrostatique du temps saisonnier (HST) est un modèle qui a été utilisé par les propriétaires de barrages pour surveiller et prévoir les déplacements d'un barrage vers un chargement externe. Cependant, le modèle HST est incapable de produire des résultats précis pour les variables qui ont des relations non linéaires, et ne tiens pas suffisamment compte des changements dus à des saisons excessivement chaudes ou froides. Dans cette thèse, trois variantes différentes du modèle HST ont été proposées pour éliminer ces défauts de l'analyse. La première variante, un modèle HST qui utilise la régression segmentée pour trouver des résultats à la relation non linéaire entre les variables. Ensuite, un modèle non paramétrique qui remplace la composante saisonnière régulière du modèle HST par une composante saisonnière non paramétrique , calculée à l'aide de la régression linéaire pondérée localement. Enfin, un modèle HST est mis en œuvre qui utilise les températures de l'air retardées autour du barrage comme détecteurs des fluctuations causées par les effets thermiques retardés. Les résultats suggèrent une amélioration de la précision de la prédiction et une réduction de la diffusion des parcelles résiduelles. De plus, deux modèles ont été proposés afin de prévoir les niveaux de piézomètre dans les extrémités amont et aval du barrage. Ces modèles utilisent des déplacements de barrage comme détecteurs supplémentaires, analysant le niveau d'eau dans le réservoir. L'auteur a conclu que la relation entre le niveau du réservoir et le

niveau du piézomètre présente une hystérésis et qu'une expression polynomiale n'est pas suffisante pour rendre compte des effets du niveau d'eau du réservoir. L'utilisation des déplacements comme variables supplémentaires améliore le R^2 ajusté globalement et réduit l'erreur standard du modèle. Un modèle segmenté a été proposé pour approcher la relation non linéaire entre les déplacements et le niveau du piézomètre qui réduit les erreurs résiduelles de près de 10% par rapport au modèle régulier.

Acknowledgements

Firstly, I would like to express my gratitude to my thesis advisor Prof. Luc. E. Chouinard, who introduced me to statistical analysis and its applications in civil engineering. It would not have been possible to complete this project without his guidance, supervision, and invaluable research advice. He consistently allowed this dissertation to be my work but steered me in the right direction whenever needed. I would also like to thank him for his financial support to help me finish the project.

I am also grateful to all other faculty members who taught me during my graduate studies at McGill University, and my undergraduate studies at Thapar University.

On a personal basis, I would like to thank my family for their consistent encouragement, moral & financial support throughout my graduate studies at McGill.

Table of contents

List of symbols

Z – standardized water level in the reservoir (reservoir level)

H – water level in the reservoir (m)

h – half width of the window

H_{min} – Minimum water level in the reservoir during the complete time series

H_{max} – Maximum water level in the reservoir during the complete time series

t – Time

T – Temperature

Dx – Displacement of the dam in the tangential axis

Dy – Displacement of the dam in the radial axis

Dz – Displacement of the dam in the vertical axis

ε – Residuals

X – Matrix of predictor variables

Y – Matrix of responses

m – Number of responses

n – Number of observations

p – Number of predictor variables

$S_{x, y}$ – Standard error of the estimate

b – vector of regression coefficients

t_o – Initial date of monitoring

t_t – Current date of monitoring

σ – variance in the residuals

α – smoothing parameter

List of figures

List of tables

1 Introduction

1.1 Brief History

Dams are one of the essential infrastructures to foster development in a country. Dams help in improving the flood control, fostering irrigation, clean energy, navigation, and recreation. Previously, dams were used mostly for irrigation in agriculture, but nowadays the primary role of dams is to generate electricity & improve flood control. Therefore, dams are a single solution to many of the critical problems faced by any developing country. Despite improved design, construction techniques, and maintenance methods, failure of dams still occur due to unforeseen loadings induced by natural forces and human actions. “Failure” in dams does not necessarily mean the collapse of the whole dam, but a collapse or movement of the part of the dam or its foundations so that the dams cannot retain the stored water (ICOLD, 1995). Dams can be classified as heavy civil infrastructures which have low frequency-high loss failure events (National Research Council, 1983). These events can be infrequent but catastrophic and can lead to significant loss of human life, infrastructure, agricultural land, and environment. Concrete dams also fail all of a sudden. The failure of a gravity dam occurs in a short period of 0.1-0.5h while the failure of a concrete arch dams may occur instantaneously-0.1h (ICOLD, 1998b). Due to this short failure time floods arising from the breach of the concrete arch dam can be more severe than embankment dams of similar heights.

Ageing of dams is another prominent concern. A large number of dams which are 30 or more years old are at a higher risk because of the increased hazard potential due to the downstream development, and increased risk due to structural deterioration or inadequate spillway capacity

(National Research Council, 1983). Dam failures in the past have led to the formation of the International Commission on Large Dams (ICOLD) in 1928. Representatives from six countries formed this commission and today constitutes representatives from more than 79 countries. The primary objective of ICOLD is to provide a platform for researchers all over the world to share their expertise and experience to alleviate problems in various aspects like design, construction, performance, monitoring, and rehabilitation of dams. For this, studies have been done to find out the major causes of failure of dams and mitigation strategies. Since its inception, the work of its technical committees and the publications of bulletins have improved dam safety, advancing state of the art safety systems.

A dam with a height of 15 meters or higher from the lowest foundation to crest, or a dam between 5 to 15 meters impounding more than 3 million cubic meters is classified as a large dam (International Commission On Large Dams, 2011). According to the world register of dams (2011), there are more than 37,640 large dams in the world, with almost half of them in China alone. With such a massive amount of aging infrastructure, improved dam monitoring systems and rehabilitation programs are necessary, since replacement of dams at such scale can be very costly for any country.

1.2 Dam Monitoring

The primary objective of dam monitoring is to extract information for assessing continued performance and safety of dams (ICOLD, 2000). Monitoring can be used for both, short term and long term. Short-term monitoring can detect short-term changes or anomalies in the behavior of the dam which can be tackled immediately by repairing or restricting any operation. Long-term monitoring detects long-term changes in the structure of the dam. It helps to predict the

deterioration in the dam and helps in choosing maintenance strategies. Analysis of the dam monitoring data utilizes techniques like multi-linear regression analysis or machine learning algorithms. These statistical models can be an assisting tool in monitoring the behavior of the dam and can assist in finding out whether the dam is behaving similarly to its past behviour under similar loads. Anomalous behavior of dam can be identified if the predicted behavior of the model does not match with the observed behavior. Visual inspections cannot help to detect any defects in the part of the dam which is submerged under water. Hence these models are used to assist dam inspections and prepare maintenance strategies.

Various types of dam monitoring systems exist, the most popular being the use of pendulums, however, alternative methods have been discussed in Chapter-3.

Dam monitoring depends on various factors like frequency of observations, placement of the pendulums, the degree of automation and type of the construction of the dam. The frequency of observations made during the construction phase and first impounding is more, as we need more information during this process to validate the assumptions in our design criteria. It is usually low during the service life as it is done to monitor long-term changes. Placement of pendulums and degree of automation are related to each other. Automated systems require less number of pendulums and give us a higher frequency of observation. Therefore they are used for short-term monitoring. Since we do not require a high frequency of observation during the service life, an automated system would not be economical. Following is a table which depicts the different criteria used for dam monitoring and their purpose.

	Category of Monitoring		
	Pre-operation	**Short-term**	**Long-term**
Aim	Knowledge of overall behavior during construction and first impounding	" Quick " assessment of safety and operability	Comprehensive assessment of safety
Number of sensors	Large	Small	Large
Reading frequency	High	High	Low
Data processing	Complex	Simple (statistics, limit values, etc.)	Simple to complex
Degree of automation	High	High	Small (partly)

Table 1.1 - Criteria for monitoring

Courtesy (ICOLD, 2000)

1.3 Motivation

The aging infrastructure and increase in the computational capacity of computers in last few decades have allowed the use of statistical analysis on large amounts of data. Historical data shows that one of the significant failures of dams is caused by loss of the stability due to uplift pressures acting on the foundation of the dam. Monitoring data from pendulums and piezometers can be helpful in analyzing the condition and reliability of such dams. However, data collection is just one step of the whole process. Analysis of the retrieved data is equally important, as it can reveal any anomalies in the behavior of the dam which can be used to alert dam wardens.

It has been noted that there is a lack of effective tools for the analysis of dam data (Dibiagio, 2000). Even though the HST model has proved to be widely accepted and accurate in estimating the displacements of the dam, it also has its own limitations. The HST model might give inaccurate estimates for seasons with high variance of temperature (Léger & Leclerc, 2007), as they do not

make use of air temperature. Also, it is based on the assumption that the Hydrostatic load and the temperature are independent of each other, however, it is well known that they are coupled as the thermal field is influenced by the water level in the upstream face (F Salazar, 2017). These defects can be removed from the HST model by adding different variables, or by adopting a different statistical technique. Also, there is a lack of models that include non-linear relationships between the reservoir level and the displacements, and the ones which also include the delayed behaviour of the dam to temperature. The inclusion of the delayed behaviour of dams will make the models more accurate, and therefore the threshold for the alarm system can be reduced further, reducing the false alarms. Even though dam monitoring has been in existence for longer than other civil engineering disciplines such as bridges and buildings, health monitoring of dams still lags behind despite improvements in sensing technologies and statistical methods. Therefore, there is a need to develop new tools to facilitate the job of dam safety engineers.

1.4 Objectives

The principal objectives of the thesis are:

a) Fitting a Hydrostatic-Seasonal-Time model to the data obtained from a dam in Quebec followed by the defects of the model. Three different HST models were made to analyze the displacements of the dam in three axes, radial (y-axis), tangential (x-axis), and vertical (z-axis). The analysis has shown various instances where the HST model is insufficient in producing an accurate analysis.

b) Improving the Hydrostatic-Seasonal-Time model by proposing three variants of the HST model. These have been proposed to tackle different problems in the statistical analysis of dam monitoring data. A segmented regression model has been proposed to tackle the non-linearity between the displacements and the water level in the reservoir. A nonparametric regression seasonal model has

been proposed to reduce the defects in the regular seasonal component of the HST model. A Hydrostatic-Seasonal-Temperature-Time model tackles the problem of the delayed effect of air temperature on the displacements of dams by using lagged variables of air temperature as predictor variables.

c) Analysis of the dam piezometer data by proposing two models to forecast dam piezometer levels upstream and downstream of the dam. The relationship between the displacements in the three axes of the dam and the piezometer levels upstream and downstream have been used to improve the existing models.

2 Statistical models used for the analysis of concrete dams

2.1 Introduction

Statistical models are an essential part of dam safety systems. After visual inspection, dam monitoring data is the most reliable way to analyze the structural behavior of a dam. Predictive models assist a dam safety engineer in analyzing monitoring data, which helps in the detection of any anomaly. Automated data acquisition systems (ADAS) utilize them on a continuous basis, assisting in timely maintenance and rehabilitation of the dam (ICOLD, 2000). These models provide us an estimated response of the dam under a given load combination, which can be compared with the observed response to draw conclusions regarding the structural deterioration (Nedushan, 2002).

Dam deformation behavior models can be constructed using deterministic methods and statistical methods. A deterministic model can predict the deformation of the dam based on the data on various factors like external forces, material properties of the dam, geometry of the dam and stress-strain relationships. Whereas, a statistical model uses years of observed data and techniques like regression analysis to estimate the future deformation of the dam within a specified confidence interval.

Deterministic models like finite element models are widely used in engineering practice to predict dam response. They are based on physical laws governing the phenomenon, and have some advantages over statistical models:

1) They are useful for the design and safety assessment for the first filling in the dam, due to the absence of significant amount of monitoring data

2) They can be interpreted conveniently, provided the parameters of the model have a physical meaning (Fernando Salazar, Morán, Toledo, & Oñate, 2017). However finite element models also have their limitations: Important stability parameters like uplift pressure and leakage flow in dams cannot be predicted accurately using numerical models (Mata, 2011). Also, the stress and strain properties of a dam and foundation materials are often limited which can lead to unreliable modeling (Salazar, Morán, Toledo, & Oñate, 2017).

The above limitations of the finite element model, combined with increased computational power and the availability of monitoring data nowadays has led to the advancement of statistical models to predict dam response. Further, many researchers have developed machine learning algorithms and artificial neural networks for the analysis of dam behavior. Most of these models employ no laws of physics or material properties and are mostly based on massive amounts of past data. This chapter will discuss various models, their applications, modeling considerations and modifications made by various researchers in the past.

2.2 Data based models for the prediction of dam behavior

Various data-based models have been proposed by researchers in the past. Most of these models utilize statistical techniques like multiple linear regression (MLR) to predict the future response of the dam to a certain load condition. Dam behavior can be attributed to three parameters: 1) the hydrostatic level of water in the reservoir, 2) variation in the air and the concrete temperature due to seasonal changes, 3) irreversible changes, like creep and shrinkage, due to time (Zhao, 2003). The effect of the hydrostatic and seasonal parameters are considered to be reversible while long-term effects of creep and shrinkage cause irreversible effects. The reversible effects are considered

during the design of dams and hence are of less concern, whereas, irreversible effects can be a concern of deterioration of the dam or any other unanticipated changes.

2.2.1 Multi-linear models to predict deformations in concrete dams

There are situations in regression analysis where more than one regressor variables are needed to fit the model. A regression model which has more than one regressor variable is called a multiple regression model. The term linear is used because the regression equation is a linear function of the parameters (or coefficients), not the variables. Following are a few multiple linear models to predict the displacements in a concrete arch dam.

2.2.1.1 Hydrostatic-Seasonal-Time model (HST)

The hydrostatic-seasonal-time model was first introduced by Électricité de France (Willm & Beaujoint, 1967). According to the Stone-Weierstrass theorem (De Branges, 1959), any continuous function closed over an interval can be approximated as a polynomial or a sum of polynomials. Similarly, the HST model is a regression model which takes into account the displacements from hydrostatic level as a fourth-degree polynomial of the water level in the reservoir, seasonal changes as a Fourier transformation of sine & cosine terms of time, and irreversible changes as a third-degree polynomial of time. A general form of a response of HST model can be formulated as follows:

$$D_i(t) = H_i(Z) + S_i(T) + T_i(T) + \varepsilon_i \qquad 2.1$$

$H_i(Z)$ is the displacement component due to the hydrostatic level of the water in the reservoir which can be expressed as:

$$H_i(Z) = a_0 + a_1 * Z_i + a_2 * Z_i^2 + a_3 * Z_i^3 + a_4 * Z_i^4 \qquad 2.2$$

Where a_0, a_1, a_2, a_3, a_4 are the coefficients of the regression analysis and Z_i is the standardized level of water in the reservoir at time‘t’.

$$Z = \frac{\text{Ht} - \text{Hmin}}{\text{Hmax} - \text{Hmin}} \quad 2.3$$

H_{max}, H_{min} is the maximum and minimum level of water in the reservoir. H_t is the water level in the reservoir at time‘t’.

$S_i(T)$ is the displacement caused due to seasonal effects. This effect can be modeled using a Fourier series, which accounts for the deformations due to change in temperature caused by various seasons. It can be expressed as:

$$S_i(T) = a_5*\sin(T) + a_6*\cos(T) + a_7*\sin(T)*\cos(T) + a_8*\sin^2(T) \quad 2.4$$

Where a_5, a_6, a_7, a_8 are the coefficients of the regression analysis, $T = \frac{2.\Pi.t}{365}$, $t = t_t - t_o$ & t_t, t_o are the current and initial date of monitoring.

$T_i(T)$ is the displacement accountable for the irreversible effects caused due to creep, shrinkage or any other unanticipated changes. Historically, various researchers have implemented different expressions to model this effect by using some heuristics and trial & error procedures. Here are a few suggested examples of the modified irreversible term of the HST model:

a) (Mata et al., 2014) introduced a linear and exponential term of time:

$$T_i(T) = a_9*t + a_{10}*e^{-t} \quad 2.5$$

b) (Yu, Wu, Bao, & Zhang, 2010) used a third degree polynomial of time:

$$T_i(T) = a_9*t + a_{10}*t^2 + a_{11}*t^3 \quad 2.6$$

Where a_9, a_{10}, a_{11} are the coefficients of the regression analysis. $t = t_t - t_o$, & t_t, t_o are the current and initial date of monitoring. ε_i is the residual error.

$D_i(t)$ is the sum of the reversible and irreversible displacements, hence the displacement of the dam at time‘t’.

The coefficients can be computed using the ordinary least squares method. Therefore the best fit model will have a minimum value of the difference between the sum of the squared values of the observations and predictions.

The model has similar assumptions as used for any multilinear regression analysis:

1) The variables chosen for the model should be independent of each other and should have a linear relationship with the dependent variable.

2) The effects of the individual independent variables on the dependent variable are additive.

3) All the errors should be independent and identically distributed.

4) The mean of the error term (ε_i) should be zero. Also, the variance of the error term should be constant.

5) The errors should be normally distributed.

Appropriate tests should be conducted to confirm these assumptions before performing the regression analysis, failing which could lead to unreliable results or an improper fit.

Changes can also be made to the hydrostatic and seasonal terms by adding or removing terms from the polynomial to fit the model better. Another alternative is to use stepwise regression, which uses the t-test p-values to remove redundant variables from the model in steps.

There are, however, limitations to this model:

1) One primary assumption is that the three effects defining the model are independent, whereas in reality there might be some correlation in between them.

2) Another one would be the lack of any physical meaning to the coefficients of the regression analysis (Bonelli & Royet, 2001)

3) H-S-T model does not consider the actual air temperature, replacing it with seasonal effects. Although it does make the model more flexible, it also reduces the accuracy of the prediction during significantly cold or warm years(Gomes & Matos, 1985).

Due to all the above limitations, various researchers have modified the HST model to include different parameters. One such model is the Hydrostatic-Temperature-Time model (HTT) (Léger & Leclerc, 2007) which replaces the seasonal effect with thermal effect by utilizing the actual temperature of the dam body using thermometers on the upstream face, downstream face, and various other specified parts of the dam. The authors have concluded that the proposed frequency-domain algorithms can convert the periodic heat problem to a transient one, and improves the fit when compared to regular HST or HT_dT. The limitation of this model is the difficulty in selecting the appropriate thermometers among the ones available.

Another model is the Hydrostatic-Seasonal-Temperature-Time model (HSTT) introduced by (Penot & Fabre, 2009), in which the periodic thermal effect is corrected with the help of actual air temperature. This model was used by Électricité de France during the European heat wave in 2003 and produced better results than the HST model (Penot & Fabre, 2009). While using air temperature data, care should be taken to account for the delay between the dam deformation and the air temperature (Bonelli et al., 2001). The HSTT model does not account for water temperature, and water temperature has been found to be a vital source of the dispersion in the HSTT model (Tatin, Briffaut, Dufour, Simon, & Fabre, 2013). Moreover, water temperature also introduces a thermal gradient throughout the structure (Tatin, Briffaut, Dufour, Simon, & Fabre, 2015). Therefore, HST-Grad model was introduced by (Tatin et al., 2015) which improves the HSTT model by considering both the air temperature as well as the gradient of the temperature in the dam body.

2.2.1.2 Principal Component Analysis of concrete dams

Multivariate methods like principal component analysis (PCA) are used when there are a large number of variables associated with multiple processes. In these cases, these variables might be correlated to each other, which can increase the dimensionality of the regression matrix making it harder to interpret. Methods like PCA reduce the number of the variables by forming linear combinations of them, hence reducing the dimensionality of the matrix. Each linear combination corresponds to one principal component. The first principal component depicts the maximum variance in the model followed by the second, third and so on. Each principal component is uncorrelated with its following principal component. The coefficients of these linear combinations are found by calculating the eigenvectors of the correlation matrix of the original variables.

(Yu et al., 2010), used PCA on the data from dam monitoring devices to extract the major principal components which describe the variance in the model. This model was further applied to an actual project, and the results showed that PCA reduces data redundancy by approximately 60%. It can also separate the noise from the signal very efficiently, which reduces false alarms.

A significant issue while using the HTT model is the selection of the thermometers. (Gomes & Matos, 1985) considered only the thermometers at the center of the cantilever span, assuming it depicts the equilibrium between the temperatures left and right parts. (Mata et al., 2014) proposed a new alternative by using PCA to select the most useful thermometers, to estimate the radial displacement of a concrete arch dam. A comparative study of the models on the Alto Lindoso dam shows that the HTT model with PCA ($HT_{PCA}T$) performs much better than the HST model, especially in dams with thinner cross-sections at higher elevations. The main reason, as stated in the study, is that at higher elevations and thinner cross-sections dam behavior is more sensitive to temperature changes. Hence the $HT_{PCA}T$ model performs better than the HST.

(Nedushan, 2002) implemented PCA on the stresses, displacements, and seepage of the Idukki dam, Daniel Johnson dam, and Chute-a-Caron dam. The author concluded that PCA could efficiently compress the original data and reduce the number of variables significantly.

Principle component analysis has its disadvantages as well:

PCA is just limited to linear combinations of the variables, and if the dependency is non-linear, then it may lead to misinterpretation of the results (Fernando Salazar et al., 2017).

PCA in concrete dam analysis is used to reduce redundancy in response variables (deformations) since we have data from a large number of pendulums. Dam analysis has typically very few predictors. Therefore PCA is rarely used to reduce the explanatory variables.

2.2.2 Artificial Neural Networks (NN model)

Artificial neural network (NN model) is a simplified mathematical model of the natural neural network. They are inspired by the efficiency of the neurons in our brain. Linear models like the MLR are not suitable to reproduce the non-linear behavior, however, with NN models there are no such limitations since they allow modeling of highly complex non-linear processes (Fernando Salazar et al., 2017). A neuron is the central element of an artificial neural network. It is an operator with inputs and outputs, associated with a transfer function. These inputs, perceptrons, and outputs are interconnected using synaptic connections or weights. The following figure depicts how the information is processed in a neuron.

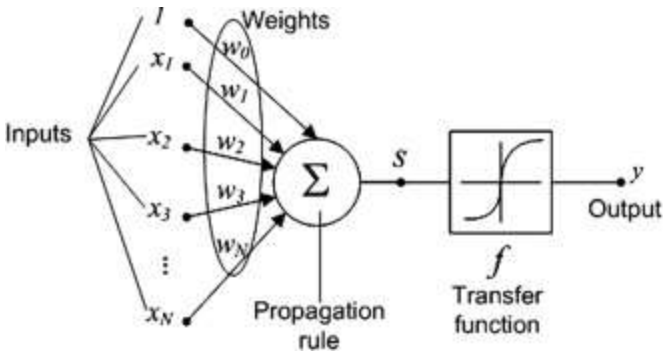

Figure 2.1 Multi-layered perception model

Figure courtesy (Mata, 2011)

There are numerous NN models available. However, multi-layer perception model (MLP) is the most widely used model in the analysis of concrete dams (Mata, 2011). MLP has neurons or perceptrons arranged in three different types of layers: input layer, hidden layer & output layer. The input layer gets the input from the data, which is then transformed by the hidden layers to form the desired output. All the neurons of each layer are connected to the neurons in the next or previous layer by synaptic connections or weights. These weights are randomly initialized at the beginning and are adjusted by an iterative process called backpropagation. When these weights are adjusted such that the desired output is obtained, we call the model to be trained. The weights of the trained model can then be used to forecast new outputs.

(Mata, 2011) did a comparative study of MLR and NN models and found out that NN model showed more flexibility and proved to perform better than the MLR model for months with extreme temperatures.

2.2.2.1 Backpropagation

A famous algorithm named backpropagation is used to train the NN model and find out the corrected weights. Backpropagation is an iterative process based on gradient descent technique, in which an input is presented to the model and output is calculated by randomly initializing the weights (Benvenuto & Piazza, 1992). Deviations between the model output and the known output are calculated, and weights are modified until these deviations are minimized. Finally, the corrected weights are saved when the deviations (cost function) is minimum or up to the desired value.

2.2.3 Model for predicting the Piezometer head in arch dams

Water seepage is a problem faced by most of the existing dams since water retained in the reservoir always find the path of least resistance. Hence it passes through the dam core and foundation and gets collected on the downstream toe of the dam. Various design considerations like impermeable core, higher compaction of the foundation, and drainage systems are considered during the design of hydraulic structures like dams, but they are only useful in controlling the amount of seepage in the dam. Uncontrolled seepage can lead to erosion of the inner surface of the dam core or foundation leading to piping. Increased uplift pressures due to soil pore pressure are also a critical part of the stability analysis of a dam.

Piezometers are instruments installed in a dam and are used to measure this pressure. Accurate forecasting of piezometer levels can be beneficial for monitoring the stability of the dam. Piezometer levels depend on the difference between the upstream and the downstream water levels. For very high dams, the fluctuations in the downstream water level are minimal compared

to the upstream water level (reservoir level) (Kalkani, 1989). Therefore, the water level in the reservoir (reservoir level) is an excellent predictor to forecast the piezometer levels in a dam.

(Kalkani, 1989) used the method of polynomial regression to fit the piezometer levels of Kremasta dam in Greece to its corresponding reservoir levels. In this method, the observations from the piezometers were fitted to a polynomial of the reservoir level using ordinary least square method.

$$\text{Piezometer level} \sim H + H^2 + H^3 + \ldots + H^n \qquad 2.7$$

H is the water level in the reservoir.

A portion of data was used to establish the model, and another portion was used as validation data for forecasting the model. The piezometer levels were separated corresponding to increasing and decreasing reservoir levels. It is mentioned that the regression curves which correspond to increasing reservoir level are lower from those which correspond to decreasing reservoir level due to a hysteresis of the total head potential at the position of the piezometer. Hence, at the same reservoir level, lower values of the piezometer levels are present when the reservoir level increases and higher values are noticed when the reservoir level decreases.

(Bonelli & Royet, 2001) also noticed a similar hysteresis of the piezometer head with increasing and decreasing reservoir levels. They suggest that the change in the piezometer level is not instantaneous with the variations in the reservoir level, but is a delayed process which can be corrected by calculating the convolution integral of the impulse response (yet to be identified) and the loading conditions (reservoir level, rainfall). The reason suggested for this delay is due to the air trapped inside the body of the dam. The authors used a Laplace transform to approximate the impulse response of the piezometer levels to the water level in the reservoir. The Laplace transformation or the convolution integral can be understood as averaging the water level in the

reservoir over time using variable weights which decay with respect to the characteristic time. The author has concluded the delayed model improves the fit of the model significantly.

3 Description of the monitoring data and dam instrumentation

3.1 Introduction

Statistical analysis is only as accurate as the quality of data recorded by the monitoring system. The main idea behind using monitoring systems and dam instrumentation is to obtain data accurately and supplementing it with visual observations to predict the structural health of the dam. A statistical model is based on various variables which are considered to be influential in predicting the response of a system. Monitoring systems furnish data for all the essential variables considered in the model. This chapter will discuss methods of data collection and the instruments used to collect data along with a few examples.

3.2 Data Collection

There are various methods of collecting data. However, each of them has their applicability. The data collected should meet the modeling objectives, which means collecting data for all the significant variables along with a specified frequency of observation. Variables that tend to change rapidly or have large standard errors should be collected at a higher frequency to reduce uncertainties. Following are the popular methods project owners use to collect data:

3.2.1 Manual data collection system

As the name suggests, manual data collection is collected manually in the field. The data is noted down in field books, tablets, paper forms, etc. Complimentary data such as date, time, instrument

name and number is usually mentioned along with the displacements. Also, the visual observations made are stored with the help of digital photographs and are saved such that it can be retrieved for future reference. Since it is a labor-intensive method, manual data collection systems are not preferred if the frequency of observations to be made is high. Therefore, they are not a good source for temporal measurements and hence can not be used for real-time monitoring systems. However, there are advantages to this system – 1) The operator can immediately check whether the data collected is anomalous and can calibrate the monitoring instrument.

2) They can prove to be a better source for spatial measurements as compared to other real-time sources like GPS or Fibre optic sensors which might be placed on strategic places on a dam (eg: construction joints, deteriorated portions etc.)

Below is a sample of manually collected data for a dam in Quebec, Canada.

Instrument name	Time	Dx (mm)	Dy (mm)	Dz (mm)	Temp (°C)	H (m)
CCCPDP6X-1	6/8/00 9:15 AM	-7.528	-3.736	-0.93	13.9	67.22
CCCPDP6X-1	2/15/99 10:16 AM	-4.22	-2.67	-1.56	16.2	67.25
CCCPDP6X-1	2/24/98 10:30 AM	-3.404	-2.57	-1.532	16	67.24
CCCPDP6X-1	10/15/97 2:30 PM	-5.445	-2.651	-0.188	15	67.25
CCCPDP6X-1	4/8/97 1:00 PM	-0.172	-0.822	-0.261	13.7	67.26
CCCPDP6X-1	3/26/96 10:00 AM	-0.375	-1.07	-0.68	14.9	67.22

Table 3.1 - Manual data

Where, Dx, Dy, and Dz are the displacements of the dam in x, y, & z-axis, and H is the height of the water in the reservoir. The figure shows that the manually collected data is obtained at an irregular frequency.

3.2.2 Automatic data collection systems (ADAS)

If a system has to be monitored at a very high frequency and real-time display & notifications are required, then automatic data collection system is a good option. It eliminates the requirement of

extensive labor by transmitting it automatically via satellite, radio or internet to a remote location, where experts can analyze the data. This way the data can be analyzed quickly and is practical during a significant event. Historically, dam monitoring systems made use of pendulums to collect the data pertaining to the displacement of the dam, however, with the emergence of new technologies remote sensing techniques make use optical and synthetic aperture radar (SAR) images of the reservoir to accurately estimate the water level in the reservoir, which is an important variable used in the analysis of dams. Global Navigation Satellite Systems (GNSS) and fibre-optic sensors help to improve the accuracy of measurements and to furnish a near real-time monitoring system. These alternate technologies became more popular following the 1971 San Fernando and 1994 Northridge earthquakes when many survey markers and access catwalks on the dam were destroyed resulting in the loss of the absolute frame of reference for the conventional surveys. GNSS, on the other hand has the absolute frame of reference far away from the site, therefore site interactions can not affect its readings. GNSS systems can capture displacements even at an interval of 10 seconds making it a near real-time monitoring system. It should be noted that this type of monitoring system is suitable only to measure static displacements, and dynamic displacements due to seismic shaking can not be measured by it since the interval is still 10 seconds.

Global Navigation Satellite Systems (GNSS) – GNSS monitoring systems are presently being used in many countries to monitor complex structural systems like dams. One such example is the integrated system that has been set to monitor the reservoir loads over three earthern dams in Hemet, USA. An active Continuosly Operating Reference Station (CORS) GNSS and a fully terrestrial geodetic system were set up to evaluate the displacements of control points located on

the crest and downstream sides of the dam within a tolerance of 10mm with a 95% confidence level. (Pipitone, Maltese, Dardanelli, Brutto, & Loggia, 2018)

Optical Fibres are also being developed to gather measurements for dam monitoring systems. The advantages of using optical fibres are many – 1) Inertness to external environment (eg: moisture, chemicals, electromagnetic fields), 2) small cross-section, 3) low signal attenuation over long distances. (Platt, Hagedorn, & Woodhead, 2011). There are three major types of fibre optic sensors that have been proposed by researchers – a) Point sensors – these type of optical fibres usually have a single sensor at the end of the fibre and the fibre itself is used to transmit the measurement.

b) Quasi-distributed – these type of optical fibres have many point sensors along the one fibre.

c) Distributed – In this type of sensor, the whole fibre acts as a sensor and measurements can be made at any point of the fibre.

Following is a table which describes the use of these types of optical fibres in dam monitoring.

"*" denotes that is relatively straightforward to adapt the sensor for this measurement

Sensor type	Bragg Grating	Interferometer	Rayleigh Scattering	Raman Scattering	Brillouin Scattering
	Quasi-distributive	Point	Distributive	Distributive	Distributive
Strain	Y	Y	Y		Y
Temperature	Y	*	Y	Y	Y
Pressure	*	*			
Displacement	*	Y			

Table 3.2 - Types of optical fibres and their uses in dam monitoring

Courtesy - (Platt et al., 2011)

Higher frequency observations help to calibrate the model better, and also reduces the errors. Even with automated systems, it is recommended to keep provisions for manual collection of data as there may be a system failure. With all the advantages, ADAS has expensive installation, potentially higher maintenance costs, and also requires a continuous source of power. Therefore most of the instruments in dams are manually operated, but a few important instruments around the construction joints, where the frequency of observation is high, are automatically monitored. Below is a sample of data collected by an ADAS system installed on a dam in Quebec, Canada.

Instrument name	Date	Temp(°C)	H(m)	Dx(mm)	Dy(mm)	Dz(mm)
CCCPDP1X-1	12/12/97 12:00 AM	-13.1	100.71	-1.08	-0.65	-1.4
CCCPDP1X-1	12/12/97 12:00 PM	-13.1	100.64	-1.08	-0.63	-1.38
CCCPDP1X-1	12/13/97 12:00 AM	-3.4	100.75	-1.1	-0.58	-1.36
CCCPDP1X-1	12/13/97 12:00 PM	-3.4	100.69	-1.14	-0.53	-1.34
CCCPDP1X-1	12/14/97 12:00 AM	-13.9	100.69	-1.18	-0.52	-1.36
CCCPDP1X-1	12/14/97 12:00 PM	-13.9	100.68	-1.18	-0.56	-1.37
CCCPDP1X-1	12/15/97 12:00 AM	-20.5	100.64	-1.15	-0.65	-1.42
CCCPDP1X-1	12/15/97 12:00 PM	-20.5	100.55	-1.14	-0.62	-1.45
CCCPDP1X-1	12/16/97 12:00 AM	-9.5	100.65	-1.14	-0.61	-1.46
CCCPDP1X-1	12/16/97 12:00 PM	-9.5	100.61	-1.17	-0.56	-1.41
CCCPDP1X-1	12/17/97 12:00 AM	-3.8	100.6	-1.2	-0.51	-1.44
CCCPDP1X-1	12/17/97 12:00 PM	-3.8	100.62	-1.24	-0.48	-1.38
CCCPDP1X-1	12/18/97 12:00 AM	-10.9	100.64	-1.27	-0.49	-1.41
CCCPDP1X-1	12/18/97 12:00 PM	-10.9	100.62	-1.28	-0.53	-1.4

Table 3.3 - Automatic data

Where, Dx, Dy, and Dz are the displacements of the dam in x, y, & z-axis, and H is the height of the water in the reservoir. Here the data is obtained at a fixed frequency of 12 hours, i.e., two observations per day.

3.3 Significant Parameters

It is vital to accurately gather data for all the significant parameters required to build the model. The automated dam monitoring systems guide by ICOLD recommends monitoring of 14 different

parameters like headwater elevation, tailwater elevation, leakage flow, drainage flow, rainfall, temperature, seismic events, phreatic surface, pore water pressure, uplift pressure, deformation, alignment & plumb, load, and total stress (ICOLD, 2000). However, the models used in this study utilize only the deformations, headwater elevation, seasonal changes (associated with time), air temperature & uplift pressures.

3.3.1 Deformations

Deformations are one of the most critical variables in the analysis. It is the dependent or response variable in the model, hence accurate data on deformations is necessary. Deformations in dams can be measured using a combination of direct and inverted pendulums. These pendulums can be used to retrieve data for both, the horizontal displacements as well as the vertical displacements. The horizontal displacements can further be divided into radial displacements (Dy) and tangential displacements (Dx).

For this study, the displacements in the z-axis were denoted by Dz and were considered the vertical displacements (along the height of the dam). The displacements in the y-axis were denoted by Dy and were considered the radial displacements (across the length of the dam). The displacements in the x-axis were denoted by Dx and were considered the tangential displacements (along the length of the dam).In a direct pendulum, the upper end of the steel wire is anchored to the top of the dam. Weight is suspended at the bottom end of the pendulum in a tank filled with damping liquid to reduce displacement due to wind or any other event. The relative displacement of the wire from the initial vertical position gives us the displacement of that particular pendulum. These relative displacements are usually stored in a readout device which can either be downloaded later or transmitted via radio, internet, etc.

In an inverted pendulum, the lower end of the pendulum is anchored to the base of the dam and a float is attached to the top of the vertical wire. This float is placed in a water tank which keeps the vertical wire in tension. The horizontal movement of the float relative to the initial vertical position provides us the horizontal displacements. To calculate the relative movements, the vertical reference frame of the inverted pendulum should be the same as that of the direct pendulum. The reading process is usually similar to the direct pendulum with a readout device which can store and transmit data. Figure - 3.1 shows a schematic diagram depicting both, the direct pendulum as well as the inverted pendulum.

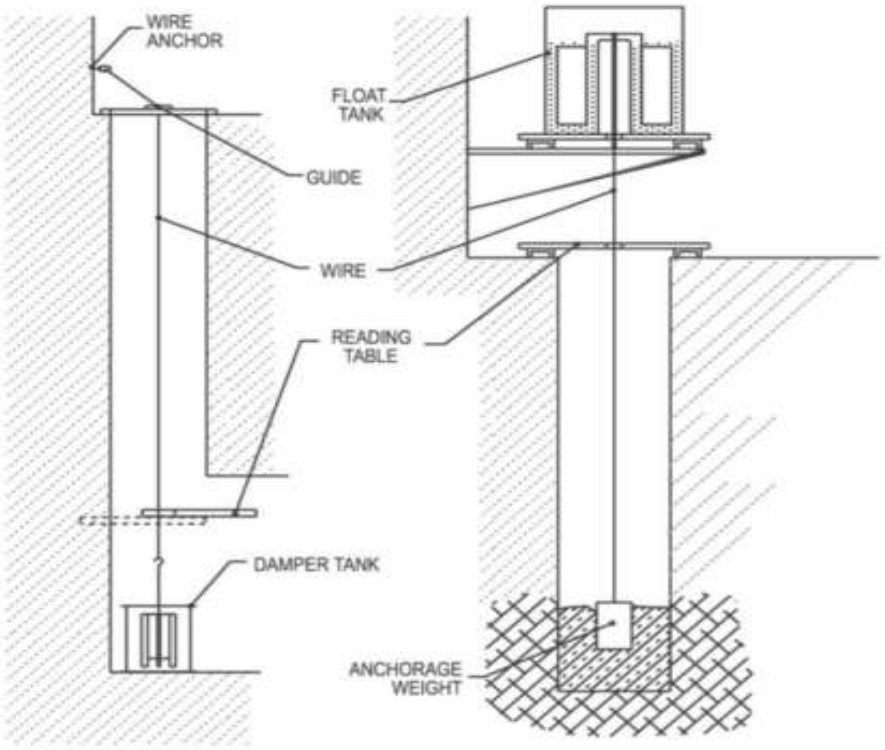

Figure 3.1 - a) Direct Pendulum, b) Inverted pendulum

Picture Courtesy (Nedushan, 2002)

In a typical dam installation, these pendulums are aligned vertically one after the other, and each pendulum is used to calculate the relative displacement between its both ends and the initial

vertical plumb. Direct pendulums are installed at the top, and inverted pendulums are installed at the bottom.

Figure 3.2 illustrates the installation of pendulums in the cross-section of a dam. Direct Pendulum-P1 calculates the displacement between points A and B, direct pendulum-P2 calculates the displacement between points B and C and inverted pendulum-P3 calculates the displacement between C and D. The inverted pendulum is anchored in firm strata and is considered fixed at the bottom.

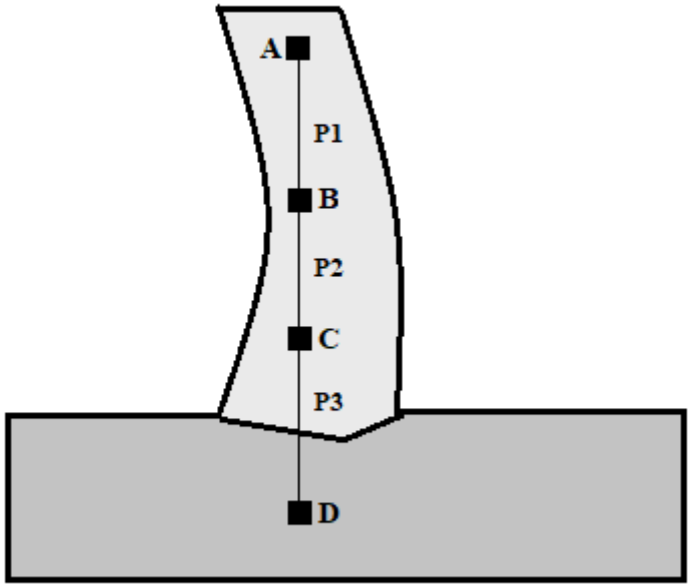

Figure 3.2 - Typical arrangement of pendulums in an arch dam

Figures 3.3, 3.4, & 3.5 show a sample output of the displacements of a pendulum as a function of time in x, y, & z-axis. Data for the deformations is available from 22nd May 1997 to 31st August 2000 with a frequency of observation of 12 hours, which gives us two observations per day.

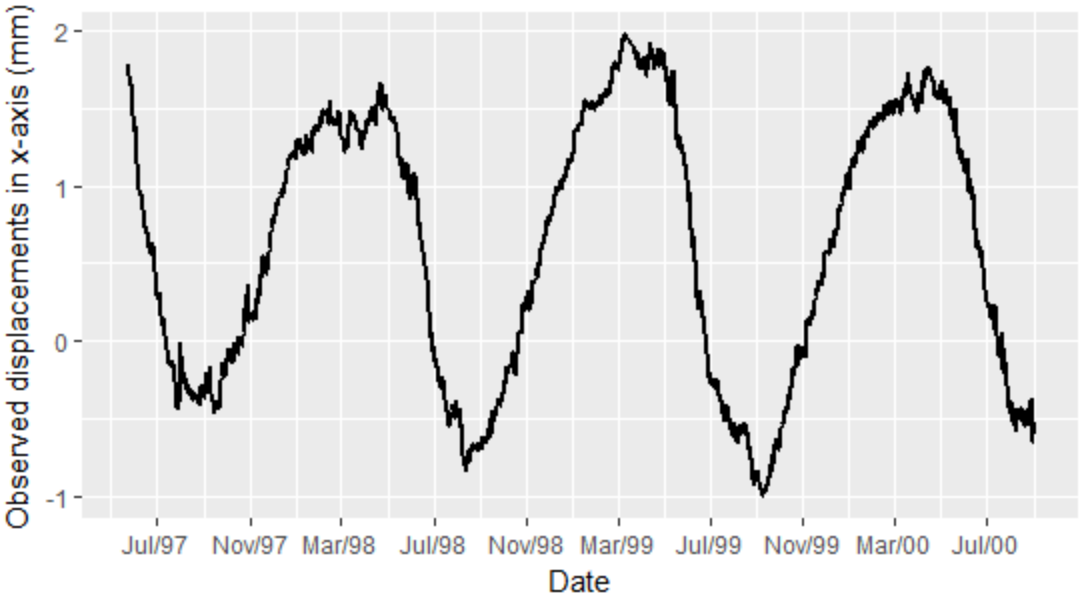

Figure 3.3 - Tangential displacement (in x-axis) as a function of time

The positive values of the tangential displacements represent the outward movement of the dam in the x-axis, whereas the negative values of the tangential displacements represent the inward movements in the x-axis.

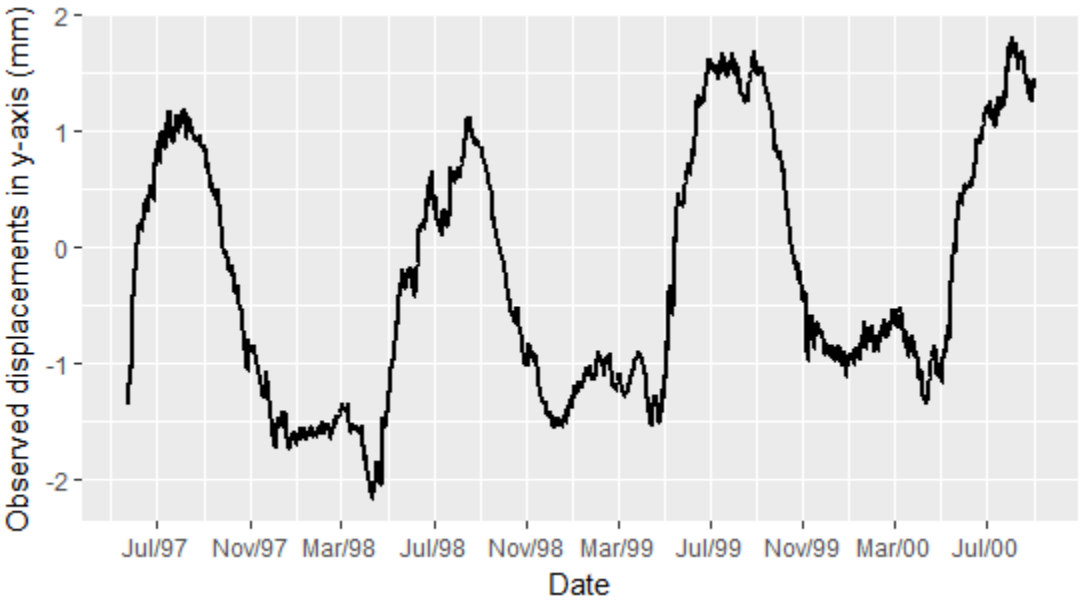

Figure 3.4 - Radial displacement (in y-axis) as a function of time

For radial displacements, the positive values explain upstream displacements, whereas negative values explain downstream displacements.

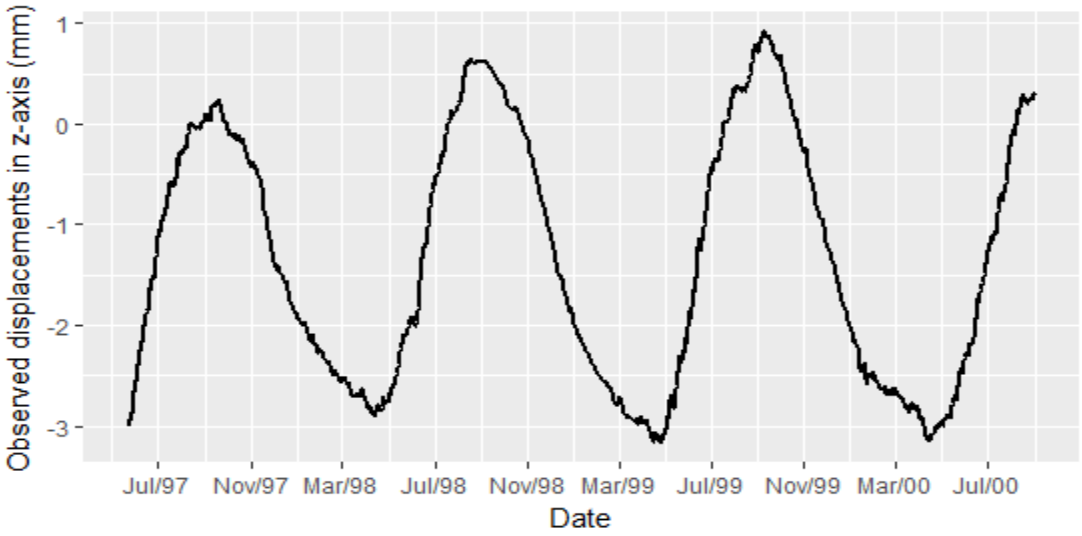

Figure 3.5 - Vertical displacement (in z-axis) as a function of time

Similarly, in the vertical axis, the positive displacements depicts increasing height, whereas negative displacements depict decreasing height.

3.3.2 Reservoir level

The reservoir level is one of the most critical predictor variables to predict the deformations in a dam. In this study, the reservoir variations follow a 12-month cycle with the maximum value as 101.94 m and the minimum value as 93.67 m. Each year the reservoir level is at its minimum during April. Figure 3.6 shows the variation of the reservoir level as a function of time.

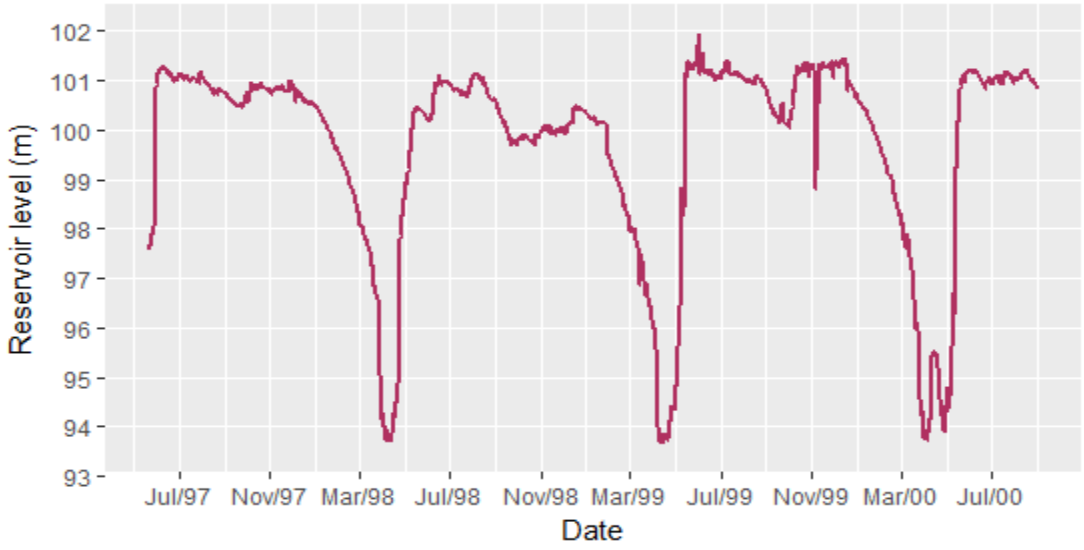

Figure 3.6 - Reservoir level as a function of time

3.3.3 Uplift pressure

Uplift pressure is crucial in defining the stability of the foundation of a dam and therefore should be monitored. It is measured with the help of piezometers which are inserted in holes drilled in the foundation of the dam. Any sudden increase in the uplift pressure is associated with instability and therefore a drainage system is provided in the foundation of a dam to reduce the pressure. In this study, two piezometers were selected, one at the downstream end of the dam, and the other at the upstream end. The piezometer at the upstream end is denoted as PZP5m, and the one downstream is denoted as PZP5v.

Figure 3.7 & 3.8 shows the variation of the uplift pressure head as a function of time in PZP5m and PZP5v. The uplift pressure is highly correlated with the reservoir level since both fluctuate in similar cycles of 12 months and reach a minimum value during April of each year. Figure 3.9 shows the typical arrangement of piezometers in a concrete arch dam

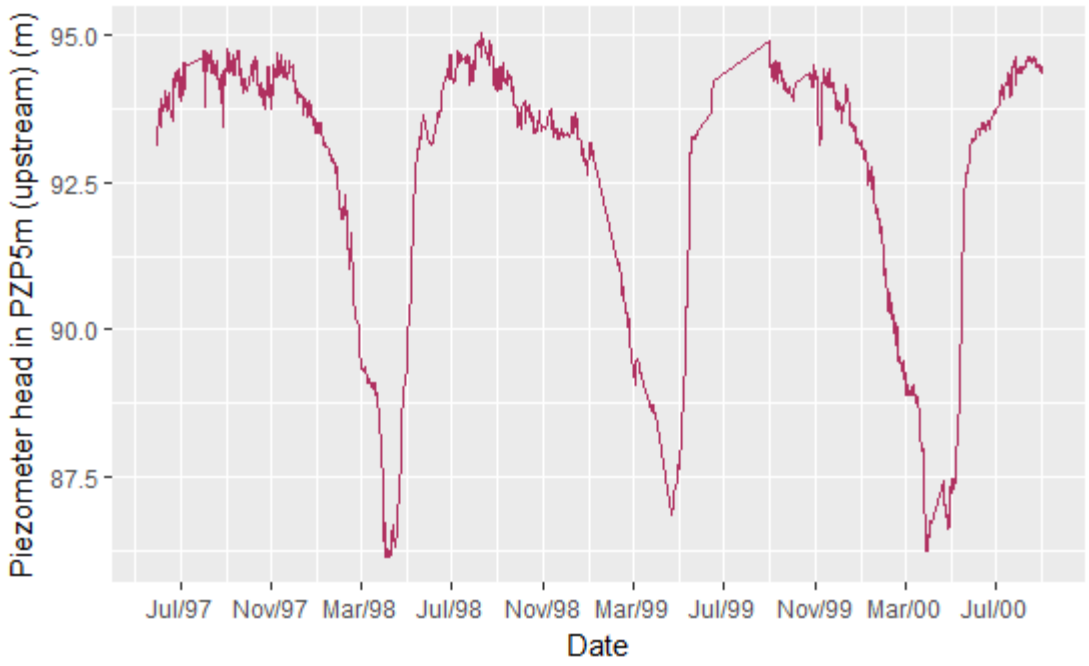

Figure 3.7 - Variation in the piezometer levels in the upstream piezometer as a function of time

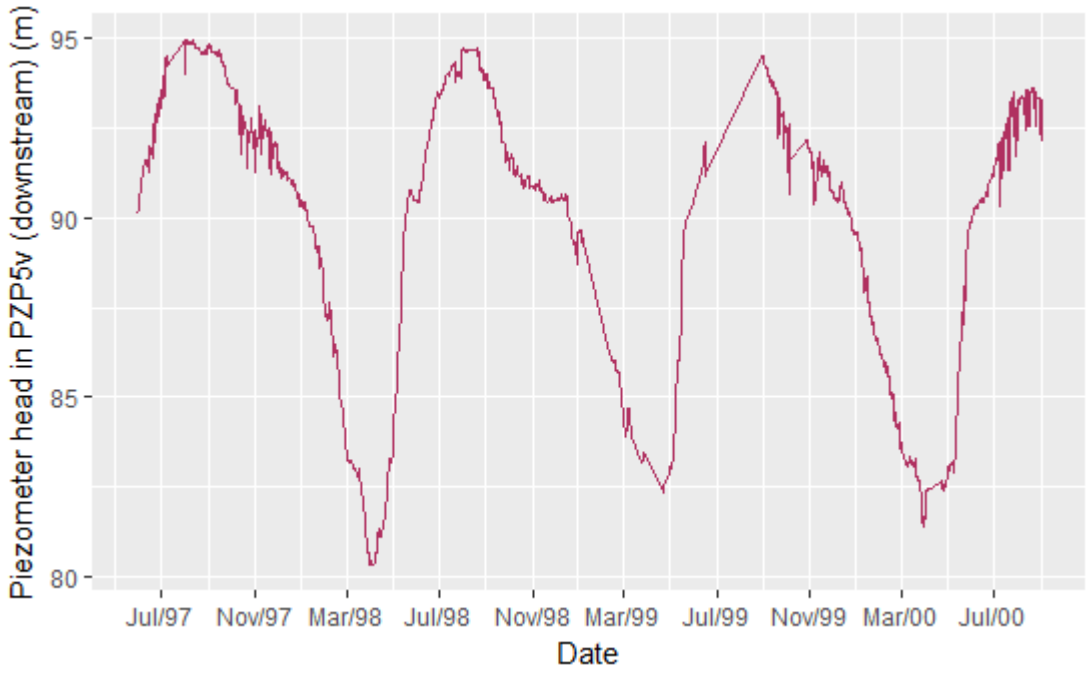

Figure 3.8 - Variation in the piezometer levels in the downstream piezometer as a function of time

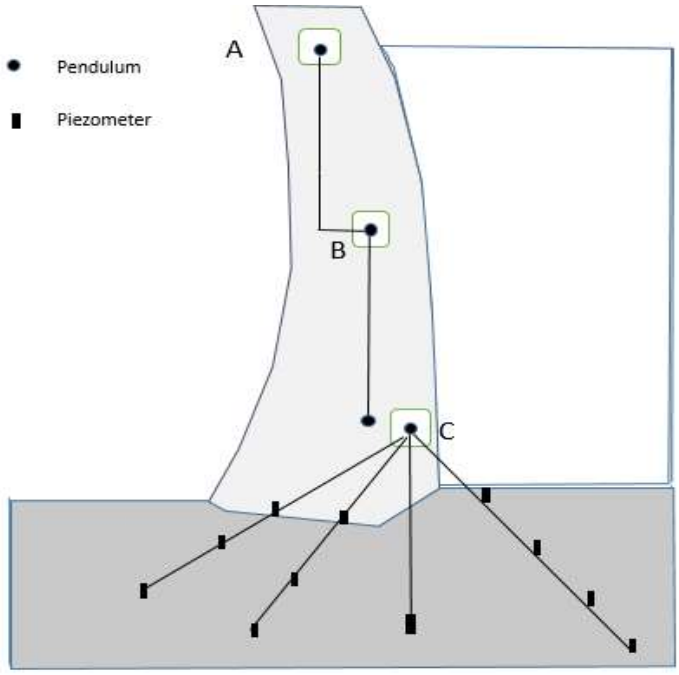

Figure 3.9 - Variation in the piezometer levels in the downstream piezometer as a function of time

3.3.4 Air temperature

The air temperature data was recorded from May 16th, 1997 to August 31st, 2000. Figure 3.10 shows the variations of temperature as a function of time. The temperature fluctuates between a max of 25.9°C to -29.7°C. The time series also shows that the seasonal variations are fairly predictable and can be used in the model instead of air temperature. Also, if the air temperature variable is considered in the model, then the temperature values would have to be smooth since the noise, and the variability in the daily air temperature data is very high

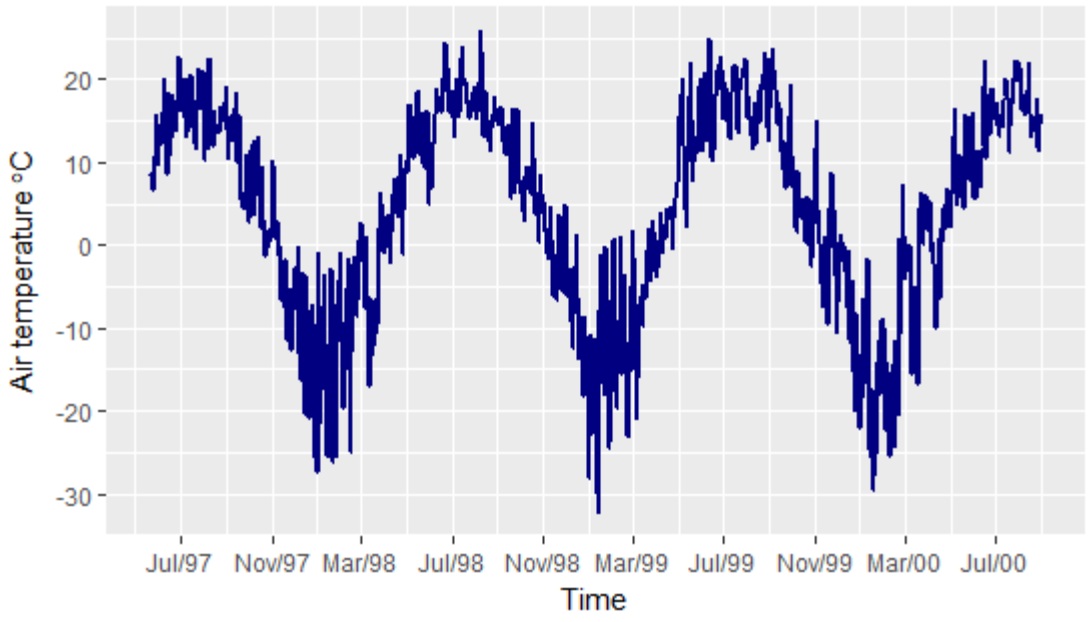

Figure 3.10 - Air temperature as a function of time

4 Methodology and a comparative study of models to predict displacement in a concrete arch dam in Quebec

4.1 Introduction

This chapter will discuss the methodology and various considerations used to build the HST model, followed by a comparative study of its different variants. The data used in this study was acquired from a concrete arch dam in Quebec. This study compares four different models to estimate and predict the deformation of a dam in three axes (radial, tangential & vertical).

4.2 Methodology

4.2.1 Elimination of outliers and smoothing of data

Monitoring data is susceptible to be accompanied by some outliers due to mistakes made during manual collection, or due to improper calibration of the instruments. These outliers can influence the coefficients produced by regression analysis, leading to a bias in the forecasting model. Therefore it is essential to remove them, to fit the model accurately. Outliers or measurement errors are mostly of two types: a) Systematic errors b) Random errors

Systematic errors are described as consistently reproducible anomalies in data, which either under predicts or overpredicts the quantity measured. These errors occur due to a constant offset in the readings, in which the instrument does not read zero when the quantity measured is zero, or due to a multiplier, in which the difference between the readings is more or less than the actual change. Systematic errors occur due to miscalibrated instruments and cannot be removed by filtering or increasing the number of observations since it is in the system itself. It can be removed or avoided by proper calibration and maintenance of the instruments.

Random errors are described as the statistical fluctuations that are random in sign & magnitude and are a result of the tolerance of an instrument. It is the inability to get the same reading every time even if the actual quantity does not change. Increasing the number of observations taken for the same quantity can reduce these types of errors.

Time series scatter plots are one of the easiest ways to observe any significant outliers from a data set. Figure 4.1, 4.2, and 4.3 show the time series scatter plots for the response & predictor variables. It is evident that there are no significant outliers in the data. The reservoir level around Nov/99 seems to be an outlier but is just a very sharp change in the reservoir level with respect to time.

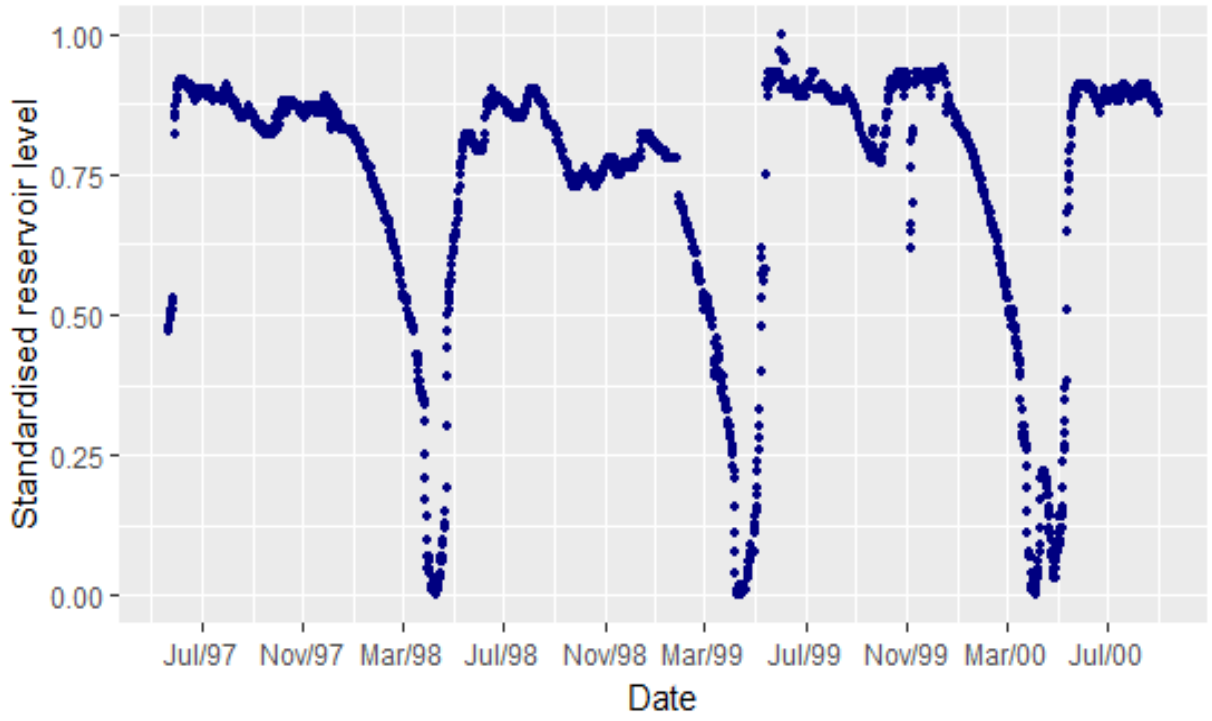

Figure 4.1 - Time series scatterplot for the standardized hydrostatic level in the reservoir

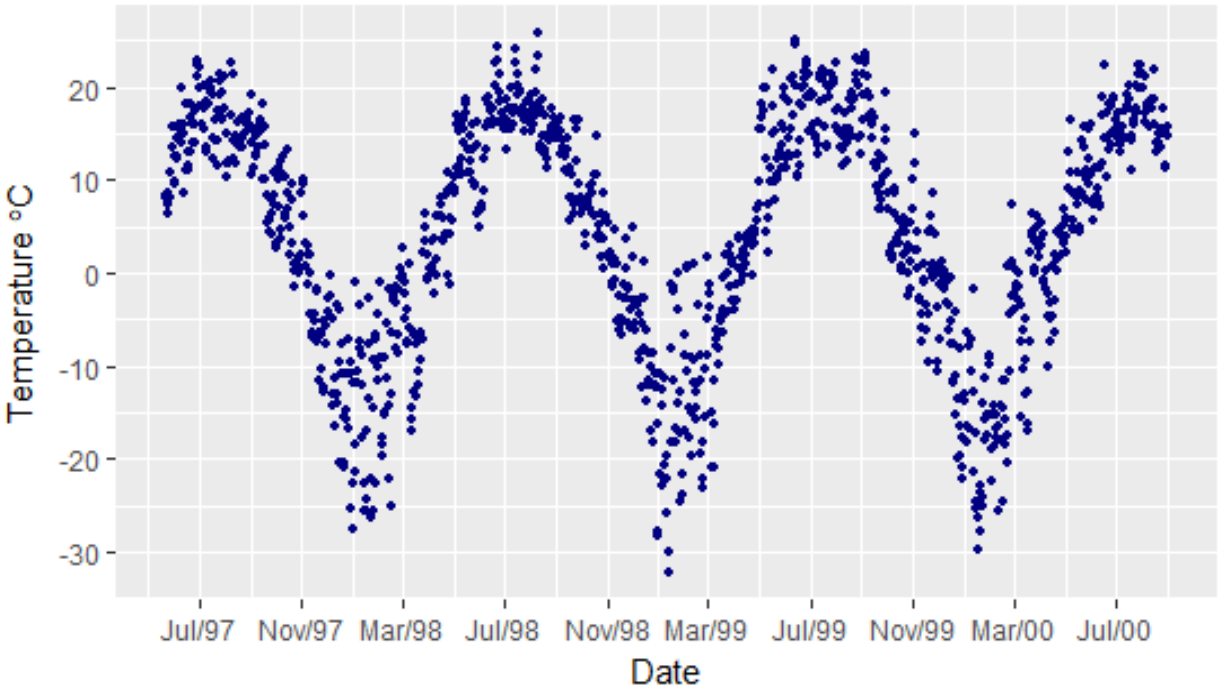

Figure 4.2 - Air temperature as a function of time

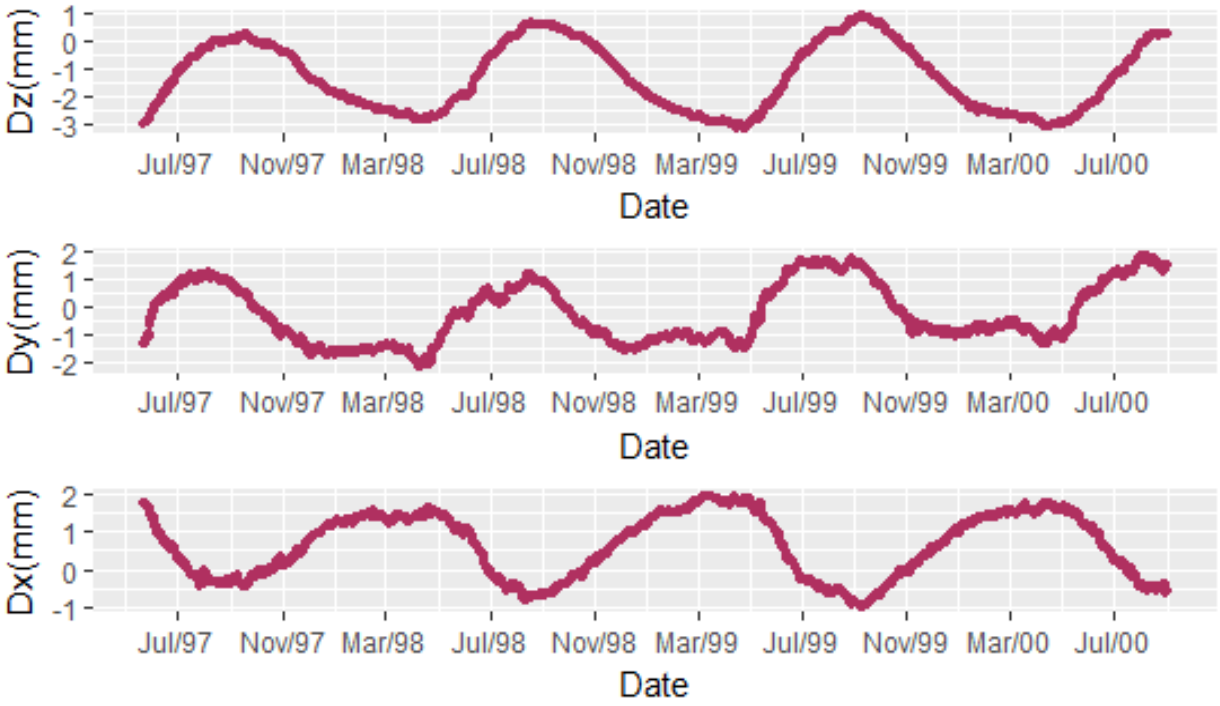

Figure 4.3 - Time series scatterplot for displacements (mm) in x, y & z-axis

In many experiments in physical sciences' & engineering, the true signal is rather a smooth curve which could be a trigonometric function or a function of time, whereas noise is seen as rapid fluctuations in the amplitude from one point to the next. Random noise should be reduced to perform the data analysis techniques accurately.

Filtering methods like the unweighted moving averages, the weighted moving averages, Savitzky-Golay filter and locally weighted linear scatterplot smoother (Lowess or Loess) aid in improving the signal to noise ratio of the dataset. Unweighted moving averages are the easiest to perform, as it replaces each value of the data set with the mean of "m" moving values in the data. Although simple, it might not be the best option since it gives equal weight to all the observations.

$$X(i) = \frac{1}{m}\sum_{i-m+1}^{i} X(i) \qquad 4.1$$

The Weighted moving average method provides a better filter since we can add higher weights to closer values and lower weights to farther values. Two popular methods are a) the triangular weighted method and b) the exponentially weighted method. As the name suggests, the triangular weighted method distributes the weights in a triangular fashion, with the closest values getting the highest weights and adjacent ones getting reduced weights, whereas the exponentially weighted method reduces the weights exponentially. These weights are decided by a coefficient α, which can vary from 0 to 1. Since α is multiplied to the nearest value and (1-α) is multiplied to the furthest values, α being closer to one means that the nearest values are more influencing than the furthest ones.

$$X(i) = \alpha * X_{i-1} + (1-\alpha) * X_i \qquad 4.2$$

Savitzky-Golay method uses least square fitting of polynomials to segments of data (Savitzky & Golay, 1964). The input variables include 1) vector of the unsmoothed signal, 2) a frame length

(should be an odd number), and 3) order of the polynomial that has to be fit. The method replaces each value of the data set with a new value obtained by fitting a polynomial of order 'k' to the adjacent values in the frame length using least square fit. Higher the difference between the frame length and the order of the polynomial, higher is the smoothing effect. Figure 4.4 depicts the comparison of two plots of air temperature data as a function of time. One of the plots is filtered using Savitzky-Golay method whereas the other plot uses no filtering.

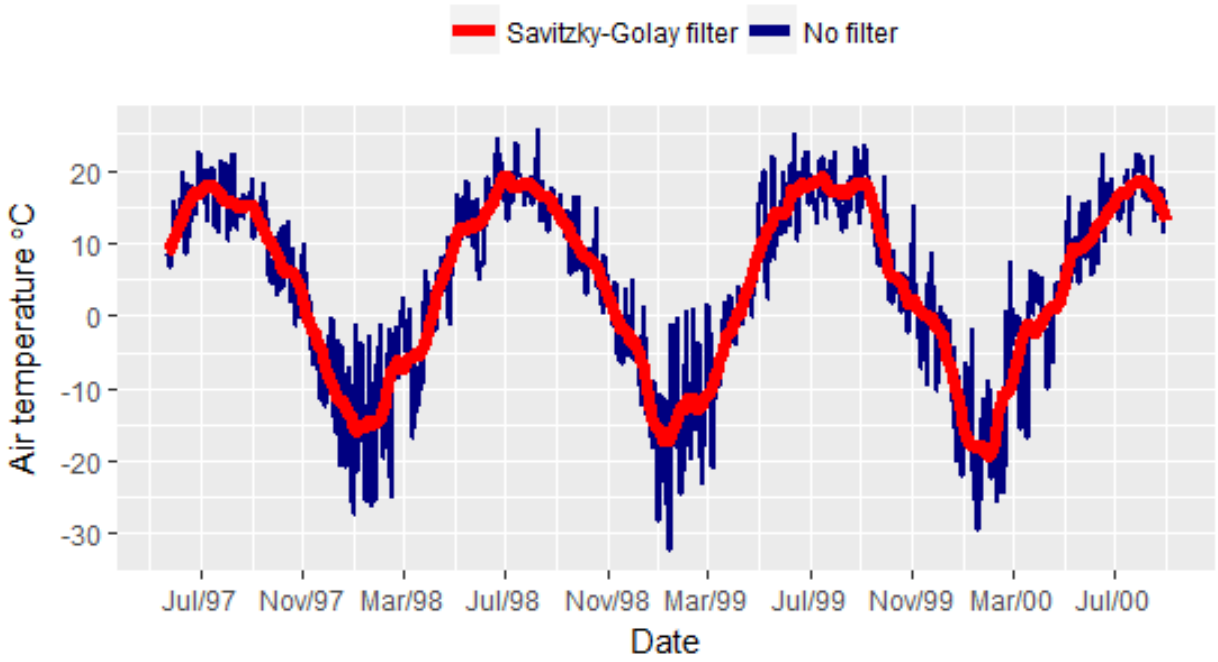

Figure 4.4 - Smoothing effect of Savitzky-Golay filter

In the figure above, a third-degree polynomial was the chosen fit for the data, with the frame size of 123 out of a total 2395 observations. The filter successfully smoothens the data values and improves the signal to noise ratio. Such smoothing filters can be applied to the data on temperature while constructing a hydrostatic temperature time model (HTT) or hydrostatic seasonal temperature time model (HSTT) since temperature data have higher noise than other variables.

Locally weighted scatterplot smoothing (Lowess or Loess) is a non-parametric filtering method pioneered by (Cleveland & Devlin, 1988) that combines the simplicity of linear least squares regression with the flexibility of non-linear regression. Loess does not require the calculation of any parameters, which further minimizes the need to make any assumptions regarding them, making it more flexible.

This method estimates the regression surface through a multivariate smoothing procedure, fitting the independent variables locally, and in

a moving fashion by choosing a specified window of a span (Cleveland & Devlin, 1988). This window or span, also called the smoothing parameter, can be varied to adjust the smoothness of the fitted curve. Larger spans result in more smoothing and vice-versa. It is called "weighted" local regression because the data points which are near to the center of the window get a higher weight in the regression analysis than the ones further. The use of a weight function is based on the idea that the nearby observations are going to be more related than the further ones. A tri-cubic function describes the weight assigned to the neighboring data points. This function is a smooth decreasing function having the maximum value at the center of the span. It is specified as:

$$W_T(z) = \begin{cases} (1-|z|^3)^3 & for\ |z| < 1 \\ 0 & for\ |z| \geq 1 \end{cases} \quad 4.3$$

Where $z_i = \left(\frac{x_i - x_o}{h}\right)$, in which x_0 is the center of the span, x_i is the i^{th} data point from the center, and "h" is half width of the span. From the equation, 4.3 one can notice that the observations further away from the half span receive zero weight from the function. Figure 4.5 depicts the distribution of the tri-cubic function over a specified span. Figure 4.6 compares the smoothing effect of Loess on the temperature data of a concrete dam with a plot of temperature with no filter. The span or smoothing parameter was chosen as 0.2, which corresponds to 480 observations out of a total of 2395 observations.

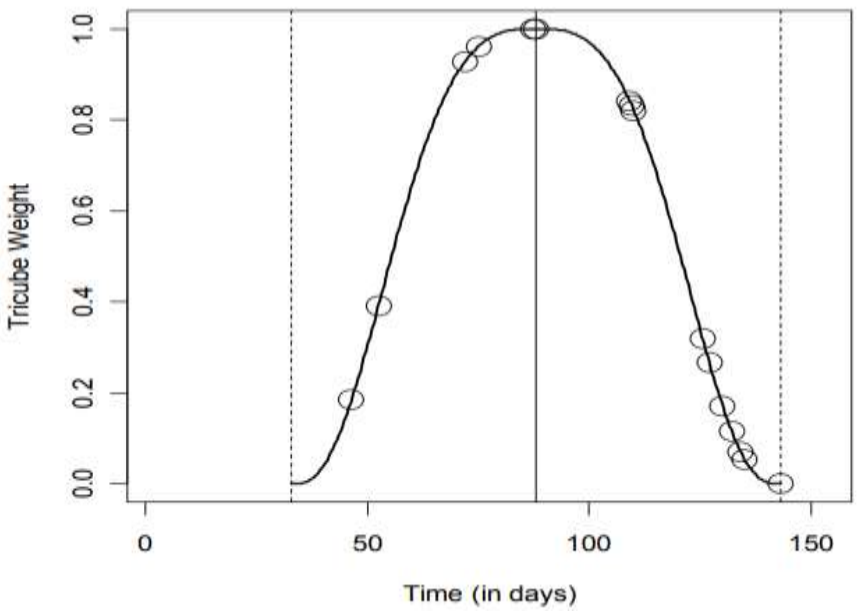

Figure 4.5 - Tri-cubic weight function

Figure courtesy (Jacoby, 2005)

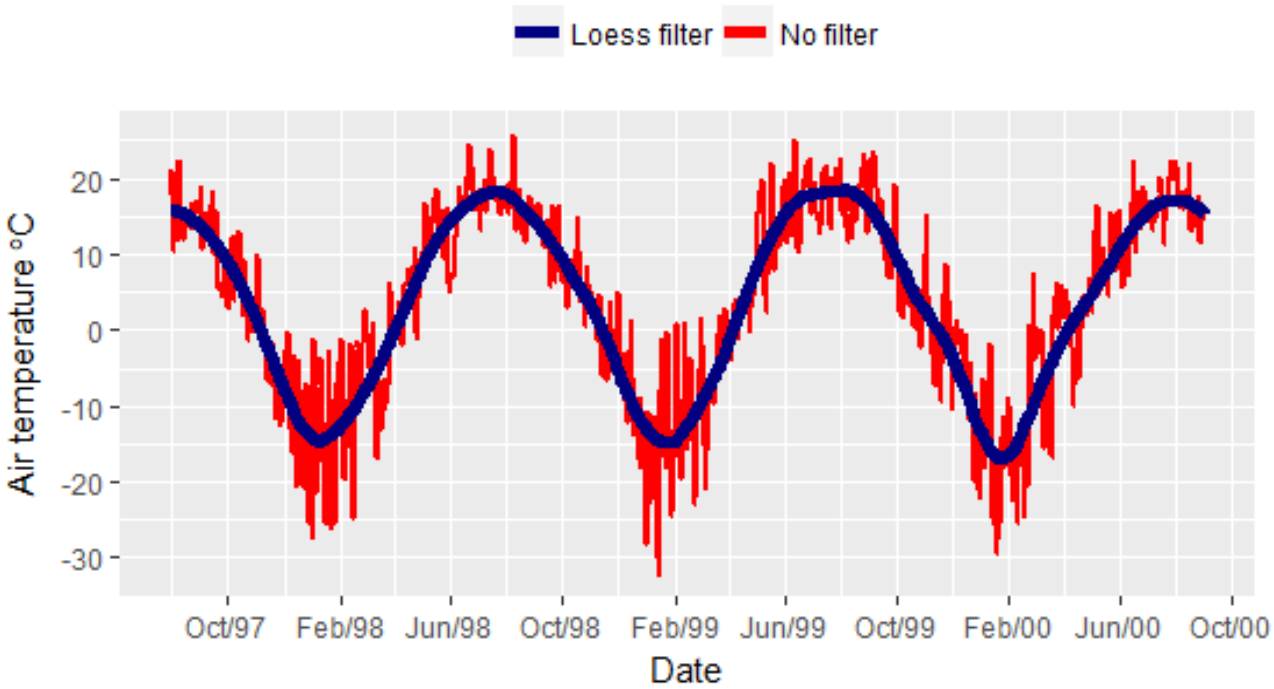

Figure 4.6 - Smoothing effect of Loess filter

4.2.2 Generalized multiple linear regression

4.2.2.1 Methodology

Multiple linear regression (MLR) is one of the most popular methods to describe relationships between variables. It is used to describe relationships between the dependent variable (response variable) Y and multiple independent variables (predictors) X_1, X_2… X_p. A generalized multiple linear regression equation looks like:

$$Y = b\,X + \varepsilon \qquad 4.4$$

Where Y is a vector of the observations of the response variable, X is the vector of the observations of the regressor (or predictor) variables, b is a vector containing the estimates of the coefficients of the regression equation, and ε is a vector of the residuals. Various methods can be used to estimate the coefficients of the regression equation, however ordinary least square estimator (OLS) was used in this study. According to this method, the estimates of the coefficients can be calculated by minimizing the sum of the square of the residuals, which means, minimizing the perpendicular distance between the observed point and the regression line.

Mathematically the square of the residuals is:

$$\varepsilon.\varepsilon' = (Y - X.b).\,(Y-X.b)' \qquad 4.5$$

Therefore, the sum of the squares of the residuals can be represented as:

$$S = \Sigma_{j=1}^{n} \varepsilon(i).\,\varepsilon(i)' \qquad 4.6$$

$$S = \Sigma_{i=1}^{n} (Y - X.b).\,(Y-X.b)' \qquad 4.7$$

The least square estimator minimizes this value using partial derivatives with respect to coefficients, to get the estimate of b:

$$\frac{\partial}{\partial b}\,[(Y - X.b).\,(Y-X.b)'] = 0 \qquad 4.8$$

$$\frac{\partial}{\partial b}[(Y.Y' - 2.Y.X.b + b'.X'.X.b)] = 0 \quad 4.9$$

Solving equation 4.9 gives us the estimate of b, $\hat{b}$

$$\hat{b} = (X'.X)^{-1}(X'Y) \quad 4.10$$

By substituting $\hat{b}$ in equation (4.4) we can find out the value of the estimate of Y:

$$\hat{Y} = X.(X'.X)^{-1}(X'.Y) \quad 4.11$$

The standard error of the estimate is expressed as:

$$S_{x,y} = \sqrt{\frac{\hat{Y}'\hat{Y} - b'X'Y'}{n-p-1}} \quad 4.12$$

Where n is the number of observations and p is the number of predictor variables

The estimates of the coefficients can then be used to forecast the response variable within a specified confidence interval. However, it should be noted that the forecasting results outside the sample data can be inaccurate.

4.2.2.2 Assumptions

Multiple linear regression is based on the following assumptions (Montgomery & Runger, 2003):

1) Linearity and additivity- The mean value of the response E(Y), at each value of a predictor X_1, X_2 .... X_n, should be a linear function of the predictors and the effect of the predictors on the expected value of the dependent variable should be additive.

2) Heteroscedasticity- The residuals should follow a normal distribution with mean equal to zero and constant variance, i.e., $\varepsilon \sim N(0, \sigma^2)$.

3) Statistical Independence - The residuals should be independent and should have no autocorrelation.

4) Stationarity – The variables, both dependent and independent, should be stationary. It means that their statistical properties like, mean, variance, and autocorrelation remain constant over time.

5) There should be no multicollinearity in the model.

Care should be taken to check whether the data set satisfies these assumptions. All of the above assumptions might never be entirely satisfied because in most of the cases there is a degree of correlation between the so-called independent variables, and this correlation leads to multicollinearity in the model. However, we must try to satisfy these assumptions to our best else the estimates of the coefficients could misinterpret the behavior of the model. Some of the assumptions like, linearity, and stationarity should be checked before running the regression analysis to avoid spurious regression coefficients.

4.2.2.3 Establishing relationships between variables

It is essential to select the correct independent variables to fit a regression model. Choosing more variables might give a higher R^2 but could lead to an overfitted model, whereas choosing fewer variables might lead to loss of information from the data. Therefore, it is recommended to visualize scatterplots and perform correlation tests before selecting the variables to be included in the model. In this study scatterplot visualization and Pearson product-moment correlation tests have been computed between the response variables (displacements) and the predictors to select correct variables and validate the assumptions of the regression model. The Pearson correlation coefficient, r, is a measure of the strength of the relationship between two variables. It can take values between -1 to +1. A value of zero indicates that there is no association between the variables and a value of +1 or -1 shows a high positive or negative association. Pearson product- moment

correlation test tries to fit a best fit line through the data between the two variables, and the value of ‘r’ shows how far the data points of the variables are from the best fit line. A value of +1 or -1 shows that all the data points lie on the best fit line, or that there are no variations from the best fit line.

Figure 4.7 shows a correlogram of the data set, which is a graph of the correlation matrix between the response and the predictor variables. It is clear from the figure that sin (T) and Z show strong associations with the displacements of the dam (Dx, Dy, & Dz). However, sin (T) has the highest correlation coefficient associated with the displacements of the dam Dx, Dy & Dz, followed by Z. It means that the seasonal components of the HST model might dominate the value of the regression coefficients more than the hydrostatic component, Z, or the time component, t.

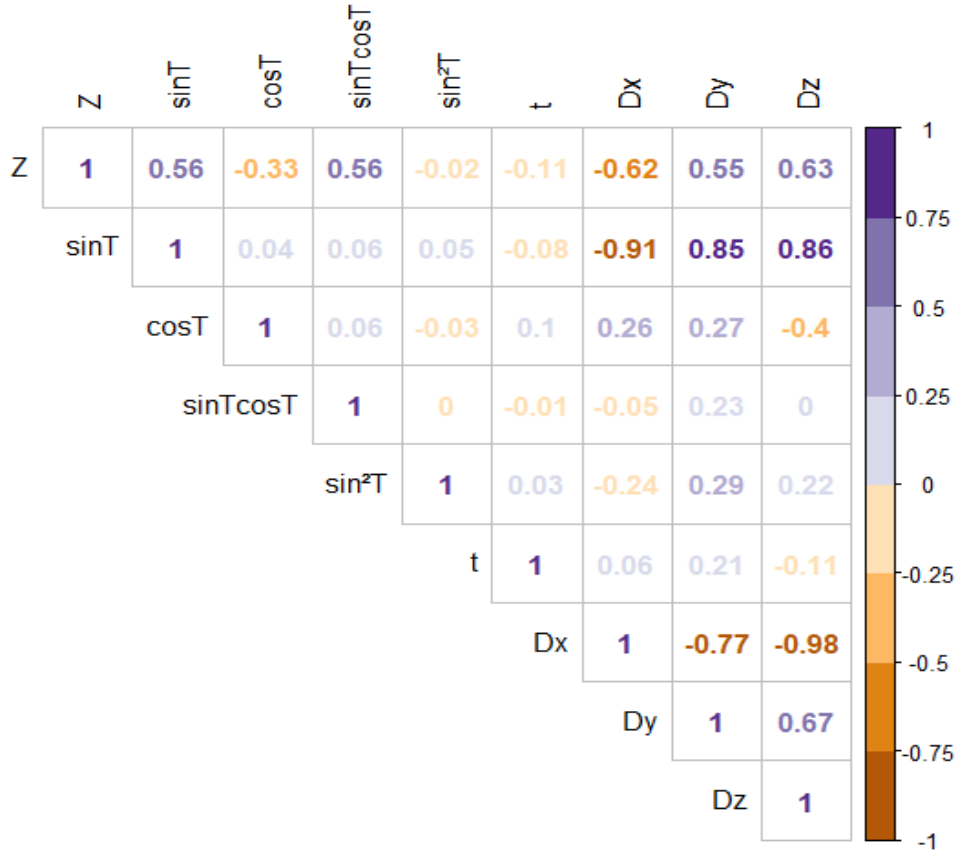

Figure 4.7 - Correlogram between the response and predictor variables

As mentioned before, one of the assumptions made for multilinear regression is linearity. Linearity means that the relationship between the dependent and independent variable is linear, keeping other variables fixed. Pearson product-moment correlation test can only compute the strength of the association between the variables but does not define whether the dependency is linear or non-linear. However, scatterplots of the variables can help confirm linearity. Figure 4.8, 4.9, and 4.10 show the nature of the relationship between the reservoir level (H) and the displacements of the dam (Dx, Dy, and Dz). The figure shows that the relationship between the reservoir level and the displacements is non-linear, hence violates the assumption of linearity. Section 4.3 of this chapter includes a piecewise regression model to tackle non-linear relationships between variables.

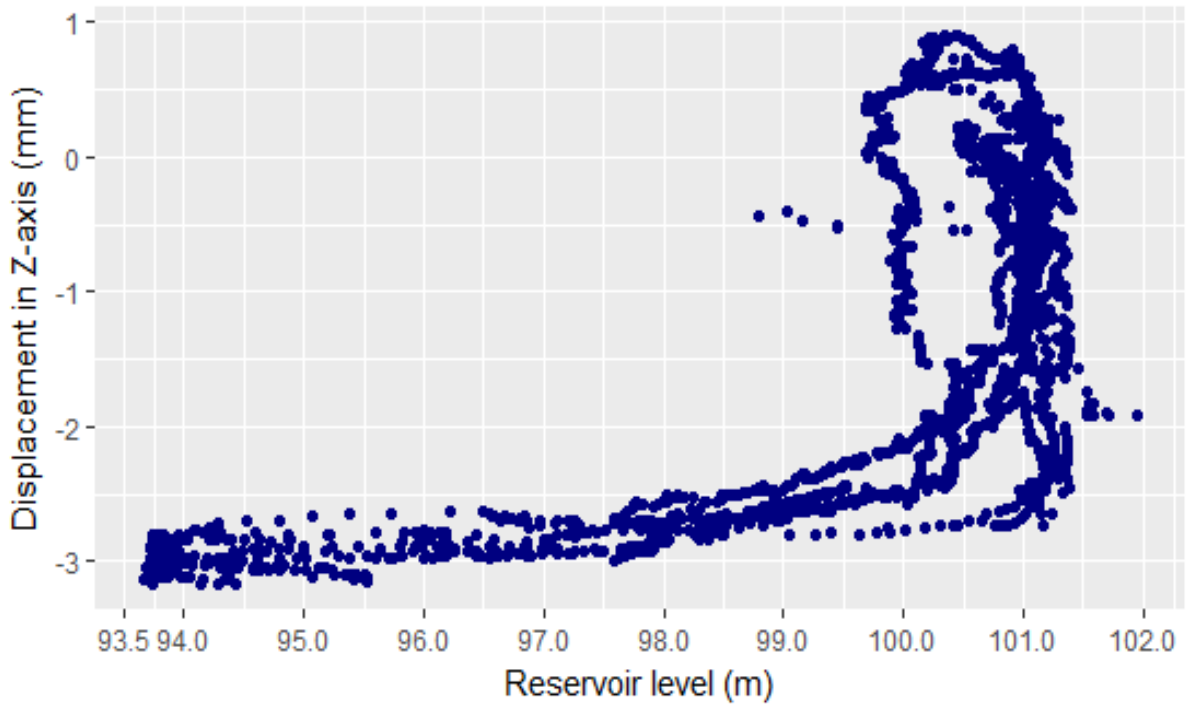

Figure 4.8 - Relationship between the reservoir level and the displacements in z-axis

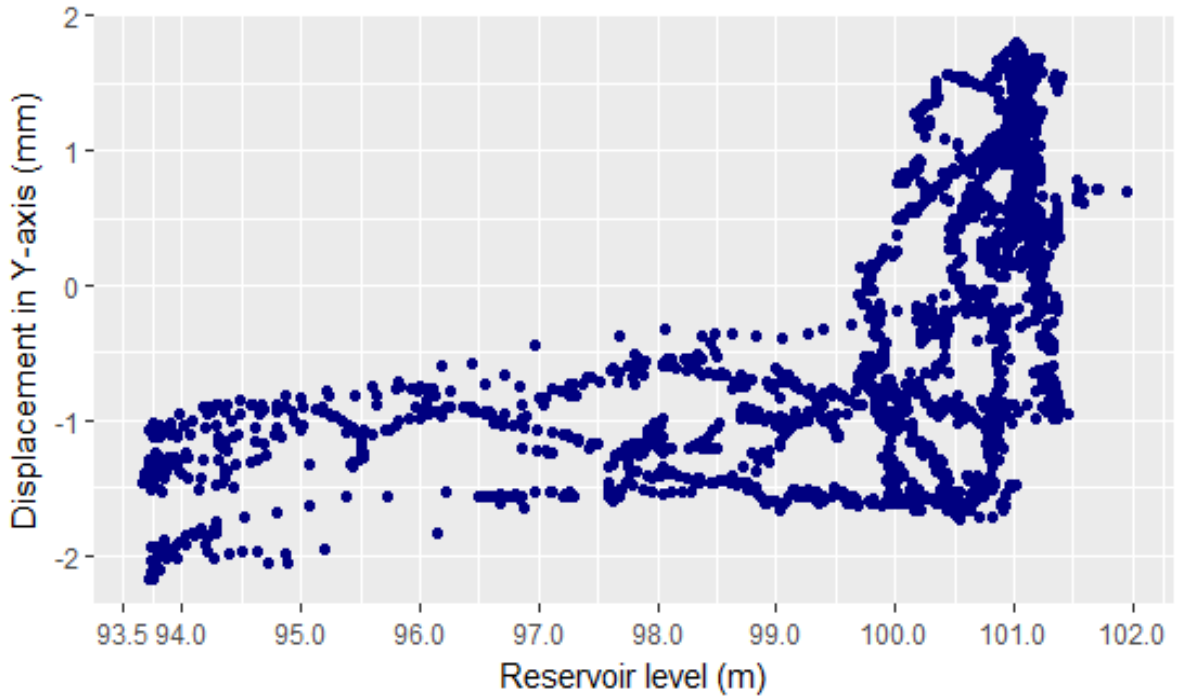

Figure 4.9 - Relationship between the reservoir level and the displacements in y-axis

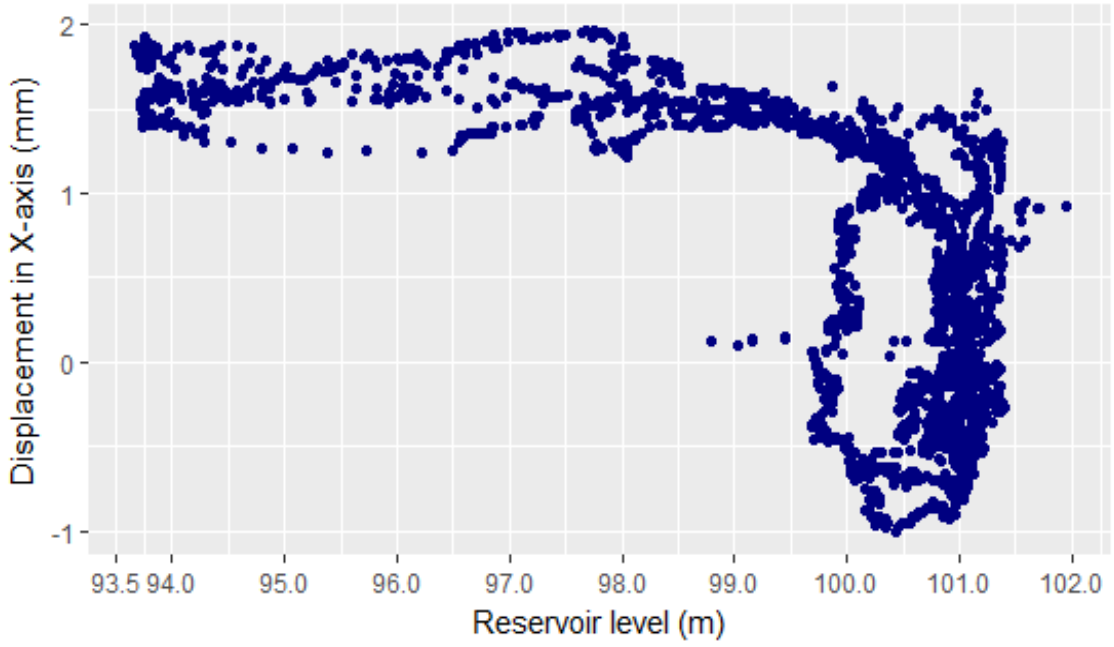

Figure 4.10 - Relationship between the reservoir level and the displacements in x-axis

It is essential to check for the stationarity of variables before proceeding with the regression analysis. Stationarity means that the statistical properties of a variable like mean, variance, and

auto-correlation remain constant throughout the time series. It implies that the distribution should have been uniform throughout the past and should continue to be uniform in future. A stationary time series is relatively easy to predict since its statistical properties will remain the same in the future as it follows a uniform distribution.

However, dam monitoring data is rarely stationary since creep, shrinkage and other irreversible displacements will lead to drift in the displacement time series. The time component of the HST model is used to capture such drift in the data but it does not account for any non-uniform variance in the displacements. This non-uniform variance is very likely if the air temperature around the dam varies a lot, or if the dam is more susceptible to daily temperature variations (thin cross-sections or high altitude). Figure 4.11 shows the time series plot for the displacement of the dam in the z-axis. It can be noticed that the amplitude of the cycles of the displacement curve for the z-axis increases over time.

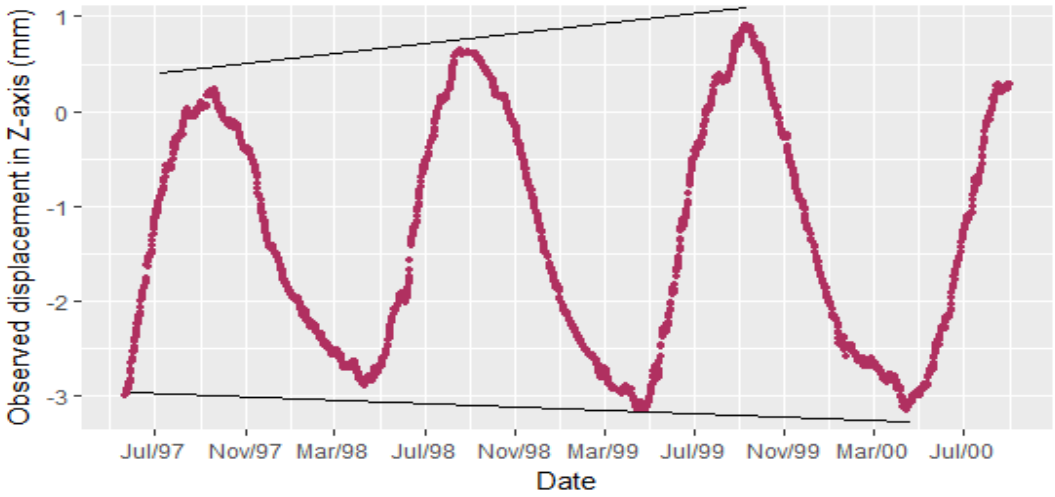

Figure 4.11 - Displacements in the z-axis over time

This increase in the amplitude shows that the data for displacement in the z-axis is non-stationary since the variance is not constant throughout. Figure 4.12 and 4.13 shows the variation of

displacement in the dam in y & x-axis respectively. The displacement in y-axis shows a drift and is also non-stationary since its mean varies over time.

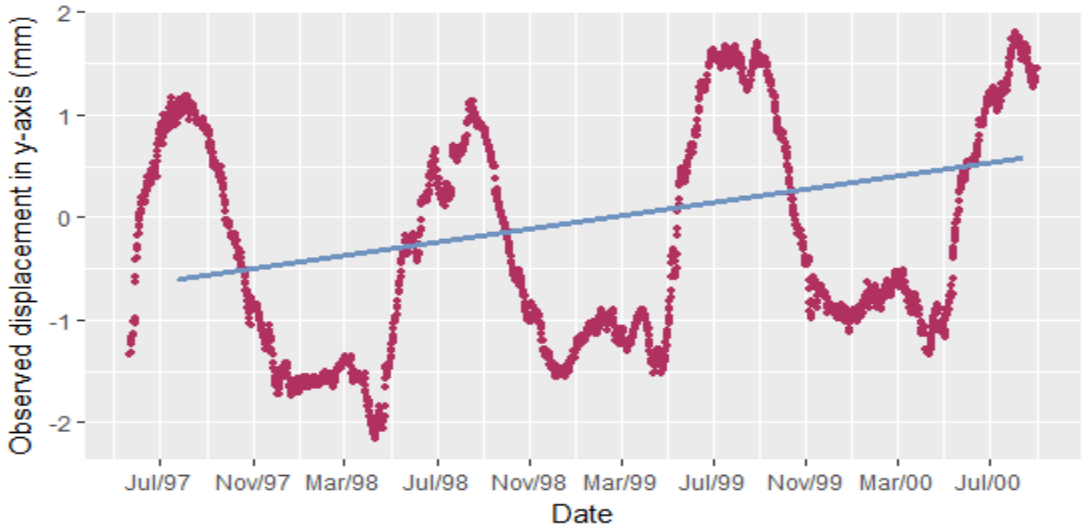

Figure 4.12 - Displacements in the y-axis over time

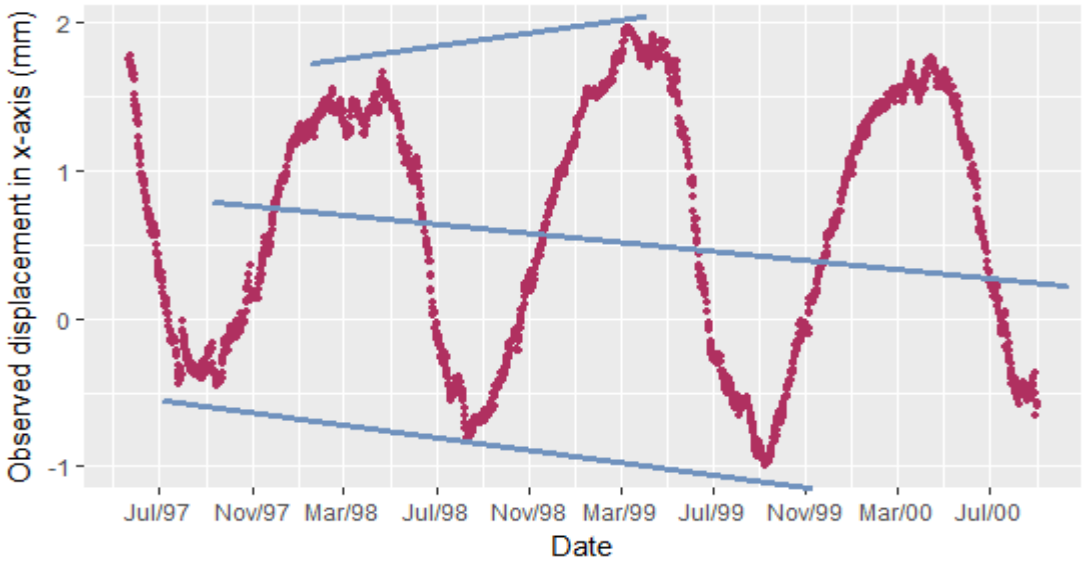

Figure 4.13 - Displacements in the x-axis over time

The displacement in x-axis shows some drift as well as non-uniform variance. Therefore, all the three displacement time series plots are non-stationary. Figure 4.14 shows the variation of the standardized reservoir level over time

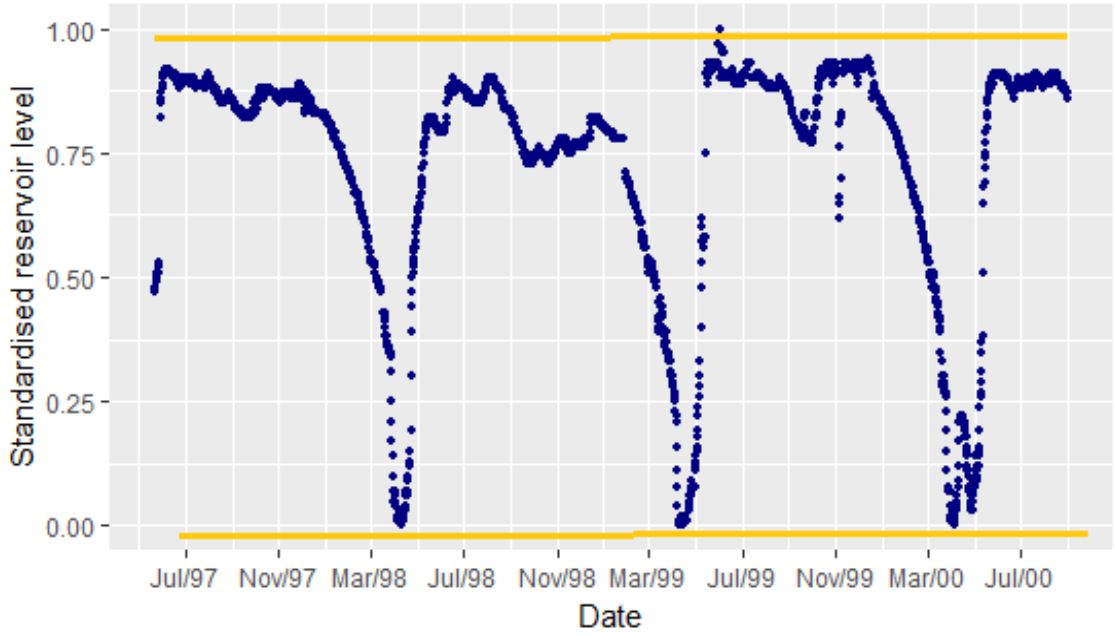

Figure 4.14 - Standardized reservoir level as a function of time

It is clear that the reservoir level curve is stationary. Therefore, the non-stationarity in the response variable has to be captured by the seasonal terms

Regression models have to be checked for multicollinearity since one of the assumptions states that there should be little to no multicollinearity in the model. A moderate to high correlation between the predictor variables lead to multicollinearity. In these cases, the calculation of the inverse of (X'X) in equation 4.9 may be difficult to obtain, since (X'X) becomes singular. (Greene, 2000) states that multicollinearity may be observed if:

a) Small changes in the data produce wide fluctuations in coefficients.

b) The coefficients may have high standard errors, and they may have the wrong sign.

Multicollinearity can be classified into two categories:

1) Structural multicollinearity: Higher orders of predictors, if included in the model, cause structural multicollinearity. E.g., If we have both x and x^2 in our model, then it is evident that they both are highly correlated. Such problem can be found in the HST model since the hydrostatic load

is defined as a fourth order polynomial of the standardized reservoir level and the time effect can also be a third order polynomial of time. Higher orders of the reservoir level were not considered in the model to reduce structural multicollinearity in the model.

2) Data based multicollinearity: In this case, the multicollinearity is caused only due to the type of data collected due to a poorly constructed experiment or due to the inability to apply any transformations to the data.

Multicollinearity must be detected and removed. Calculation of the variance inflation factor (VIF) is a useful method of detecting multicollinearity. As mentioned before that multicollinearity increases the standard errors of the coefficients, therefore, there must be an increase in the variance of the coefficient. As the name suggests, variance inflation factor quantifies such inflation in the variance of the coefficient. A VIF of one means no inflation in the variance of that particular coefficient compared to no multicollinearity, VIF of 5-10 is considered to be moderate and depicts some multicollinearity, but a VIF > 10 shows considerable multicollinearity and can influence the values of the coefficients. The variables which have a VIF>10 should be removed from the model, until all the coefficients in the model have a VIF<10, preferably below 5.

Performing stepwise regression can also reduce multicollinearity in the model. Stepwise regression utilizes a unique algorithm for selecting variables for the model. It chooses the variables by comparing their t-test p-value with the significance level (confidence level), as specified in the model. Variable with the least p-value gets selected first, and then other variables are added into the model according to their increasing p-values until the fit of the model stops improving, or until there are no other variables with p-values less than the significance level. There are three ways to run stepwise regression models:

a) Stepwise regression (forward): In this method, the first variable selected for the model is the intercept, and then other specified variables are added into the model by comparing their p-values with the significance level until all the variables are exhausted.

b) Stepwise regression (backward): In this method, all the variables are initially considered in the model, and then insignificant variables are removed from the model in steps by comparing their p-values with the significance level specified in the model.

c) Stepwise regression (both): In this method, both forward and backward regressions are performed, and insignificant variables from both are removed from the model similar to the above two methods.

Alternative methods to remove multicollinearity are multivariate methods like principal component analysis (PCA) and partial least square method (PLS), which remove the redundant variables from the model by forming linear combinations of them.

4.2.2.4 Residual analysis and validation

Residuals or errors are the difference between the observed response and the expected response which is predicted by the regression equation. In regression analysis, the error term should be random or unpredictable. We can breakdown regression equation in deterministic and stochastic parts.

Response = Deterministic + Stochastic

Deterministic part is the one which should have all the predictive power of the model. The stochastic part or error part should be random and should follow a normal distribution with zero mean and a constant variance. Residual plots are those plots where the error is plotted on the y-axis with a predictor variable or response variable on the x-axis. Visualization of the residual plots

is an important step in regression analysis since a white noise residual plot confirms that there is no predictive power left in the data and that only randomness is left behind. Residuals in the plot should be centered around zero and should have a constant variance. Any pattern or drift in the residual plot shows that some predictive term has still not been captured by the model.

Residual analysis helps us to know whether our model has all the predictive terms, or if some variable can be added to the model.

Also, the model should be validated to assess the predictive capacity of the model on new observations. There are a few methods to validate a regression model.

1) Validation set approach: In this method, the data is split into training and testing subsets. The regression model is trained using the training data, and further, the regression equation is used to predict the observations in the testing subset. The resultant validation-set error is a reasonable estimate of the test error. The disadvantage of this method of validation is that only a fraction of the data is used to fit the model and it could lead to a weaker fit if the observations are few.

2) K fold cross-validation: The data is divided into k equal parts, with k-1 parts used to fit the model and the k^{th} part is used to validate the predictions. The model trained using the k-1 parts is used to predict the k^{th} part and error is computed. This process is repeated iteratively for all values of k, and the resultant error is averaged over k. So if k=4 then the process is repeated four times with k=1, 2, 3, & 4.

3) PRESS statistic or leave one out approach: PRESS (prediction sum of squares) or leave one out approach allows the whole data to be used a training data. In this method, one data point is removed from the whole dataset upon which regression analysis is performed again to predict the observation that was left out. This method is repeated for all data points in the dataset, and the

predictive performance of the model can be checked by how well it predicts the left out observations.

Predicted R^2 is calculated from the PRESS statistic and is much better criteria to access the predictive power of the model than R^2 or Adj.R^2.

4.3 Comparative study of different models for predicting dam displacement

This comparative study has been done to assess the performance of four different variants of the HST model, to predict the displacement of a dam in three axes, radial, tangential, and vertical. Comparisons between the fit of these models have been discussed along with the regression diagnostics.

4.3.1 Model -1 (HST)

Model -1 is a hydrostatic-seasonal-time (HST) model which uses stepwise regression to fit the dam monitoring data. The HST model can be decomposed into three different terms: a) Hydrostatic deformations, b) Seasonal deformations, c) Irreversible (time) deformations.

The model is expressed as a polynomial of the standardized hydrostatic level, trigonometric functions and time. Individual models named HST_X, HST_Y, and HST_Z were made to fit the displacements in three axis, x,y, and z:

$$\text{HST\_Z} \rightarrow Dz \sim a_0 + a_1 * Z_i + a_2 * \sin(T) + a_3 * \cos(T) + a_4 * \sin(T) * \cos(T) + a_5 * \sin^2(T) + a_6 * t + a_7 * t^2 + a_8 * t^3 \quad 4.13$$

$$\text{HST\_Y} \rightarrow Dy \sim a_0 + a_1 * Z_i + a_2 * \sin(T) + a_3 * \cos(T) + a_4 * \sin(T) * \cos(T) + a_5 * \sin^2(T) + a_6 * t + a_7 * t^2 + a_8 * t^3 \quad 4.14$$

$$\text{HST\_X} \rightarrow Dx \sim a_0 + a_1 * Z_i + a_2 * \sin(T) + a_3 * \cos(T) + a_4 * \sin(T) * \cos(T) + a_5 * \sin^2(T) + a_6 * t + a_7 * t^2 + a_8 * t^3 \quad 4.15$$

Dx, Dy, Dz represent the modelled displacements (in mm) of the dam in tangential, radial, & vertical axis respectively. The higher orders of the hydrostatic terms were removed to reduce structural multicollinearity in the model. A stepwise regression algorithm was used to find out the coefficients for the above three equations, making them deterministic. These equations can then be used to forecast the displacments for the three axes. Tables 4.1, 4.2, & 4.3 show the regression summaries of the stepwise regression analysis for the three axes:

Independent variables	Coeff.Estimate	Std.Error	t-value	p-value(>|t|)
Intercept	-2.227e+00	2.495e-02	-89.228	< 2e-16
Z	1.463e-01	2.694e-02	5.429	6.23e-08
sin (T)	1.543e+00	6.915e-03	223.206	< 2e-16
cos (T)	-7.059e-01	5.978e-03	-118.076	< 2e-16
sin(T).cos(T)	-7.150e-02	1.387e-02	-5.154	2.76e-07
$\sin^2$ (T)	5.734e-01	9.309e-03	61.598	< 2e-16
t	2.508e-03	9.603e-05	26.113	< 2e-16
t^2	-2.897e-06	1.848e-07	-15.675	< 2e-16
t^3	7.641e-10	1.012e-10	7.552	6.05e-14
Residual standard error			0.1597 on 2386 Degree of freedom	
Multiple R-squared			0.9833	
Adjusted R-squared			0.9832	
Predicted R-squared			0.9831	

Table 4.1- Regression Summary for model-1 in the vertical axis (z-axis)

Independent variables	Coeff.Estimate	Std.Error	t-value	p-value(>|t|)
Intercept	-1.790e+00	2.272e-02	-78.79	<2e-16
Z	1.340e+00	2.022e-02	66.26	<2e-16
sin (T)	9.881e-01	6.530e-03	151.32	<2e-16
cos (T)	5.008e-01	5.867e-03	85.36	<2e-16
$\sin^2$ (T)	8.235e-01	1.029e-02	80.04	<2e-16
t	-3.130e-03	1.064e-04	-29.40	<2e-16
t^2	7.997e-06	2.048e-07	39.04	<2e-16
t^3	-4.396e-09	1.121e-10	-39.22	<2e-16
Residual standard error			0.177 on 2387 degrees of freedom	
Multiple R-squared			0.9721	
Adjusted R-squared			0.9720	
Predicted R-squared			0.9718	

Table 4.2 - Regression Summary for model-1 in the radial axis (y-axis)

Independent variables	Coeff.Estimate	Std.Error	t-value	p-value(>|t|)
Intercept	1.129e+00	2.382e-02	47.415	< 2e-16
Z	-3.948e-02	2.780e-02	-1.420	0.156
sin (T)	-1.107e+00	7.146e-03	-154.950	< 2e-16
cos (T)	3.383e-01	6.179e-03	54.739	< 2e-16
sin (T).cos(T)	-5.871e-02	1.433e-02	-4.096	4.34e-05
$\sin^2$ (T)	-4.430e-01	9.593e-03	-46.181	< 2e-16
t	-8.635e-04	4.098e-05	-21.072	< 2e-16
t^2	6.381e-07	3.317e-08	19.240	< 2e-16
Residual standard error			0.165 on 2387 degrees of freedom	
Multiple R-squared			0.962	
Adjusted R-squared			0.9619	
Predicted R-squared			0.9617	

Table 4.3 - Regression Summary for model-1 in the tangential axis (x-axis)

The seasonal terms like sin (T), cos (T) were found to have high t-values and low p-values in all three axes. The t-values were found to exceed 100 in all three axes which means that the seasonal component of the model-1 dominates the calculation of the regression coefficients.

Another important observation was that the influence of the hydrostatic term, Z, was most significant in the radial axis (along the water body) with a t-value of 66. It makes sense theoretically since the change in reservoir level should produce a higher effect in the radial displacements than the tangential and vertical, since it is across the length of the dam. Similarly, the term Z was the least influential in the tangential axis (-1.42) since it is perpendicular to the radial axis and changes in reservoir level should not produce much effect. The adjusted R–squared statistic was found to be high (>0.96) and the root mean square error was low (<0.177) for all three axes. The predicted R-squared calculated by using the PRESS statistic using leave-one-out approach was also high (>0.96). Figures 4.15, 4.16, & 4.17 compare the observed displacements of the dam with the predicted displacements and depict the decomposition of the displacements in hydrostatic, seasonal, and irreversible terms in x, y, and z-axis.

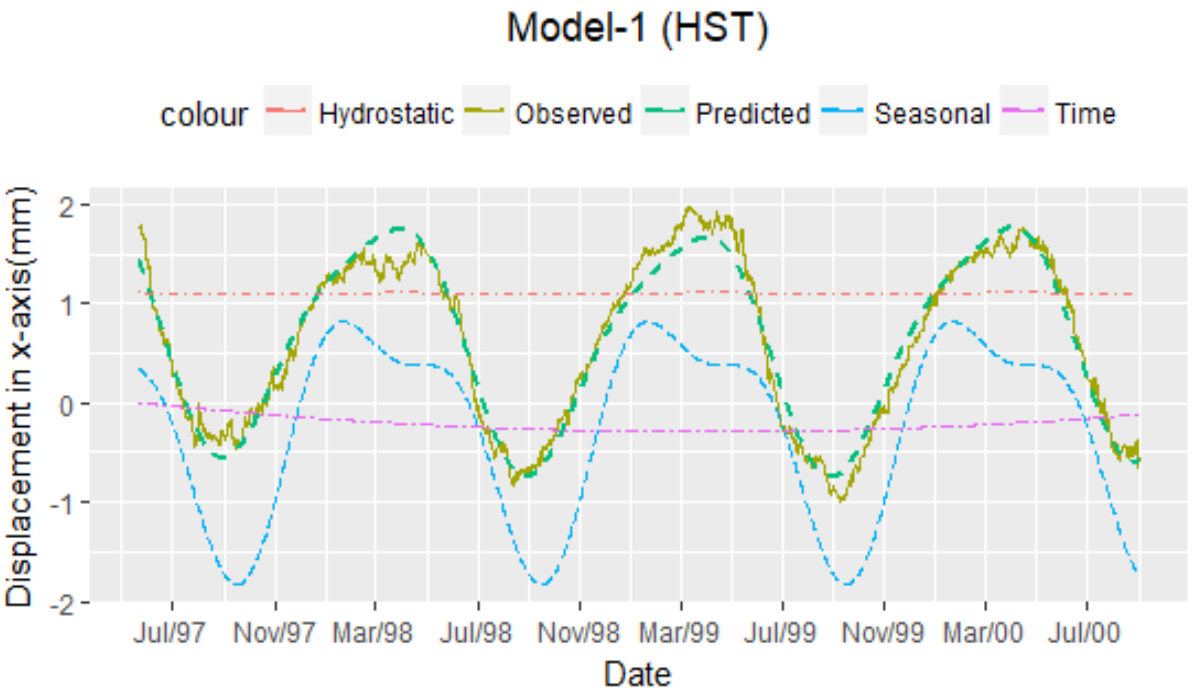

Figure 4.15 - Comparison between the predicted and the observed displacements in the x-axis

The above figure shows the comparison between the predicted displacements and the observed displacements in the tangential axis, with a decomposition of the predicted displacements into three components of the HST model. It can be seen that the prediction of displacements is accurate except for the months of March & April. The improper fit during these months is because the air temperature data (Figure-4.2) is highly variable during January to March, and since dam response is delayed with respect to the air temperature, the fluctuations in the displacements are seen in between March &April.

It can also be noticed that displacements in March/2000 are less variable than the previous years since the temperature during January to March in the year 2000 is less variable than its previous years. Therefore it can be said that these deviations between the predicted and the observed displacements are because the seasonal term of the HST model can not account for the daily

temperature variability, as it is based on smooth trigonometric functions of time. Also, since the displacement in the x-axis is more influenced due to the seasonal term than the hydrostatic term; it leads to substantial residuals during those two months.

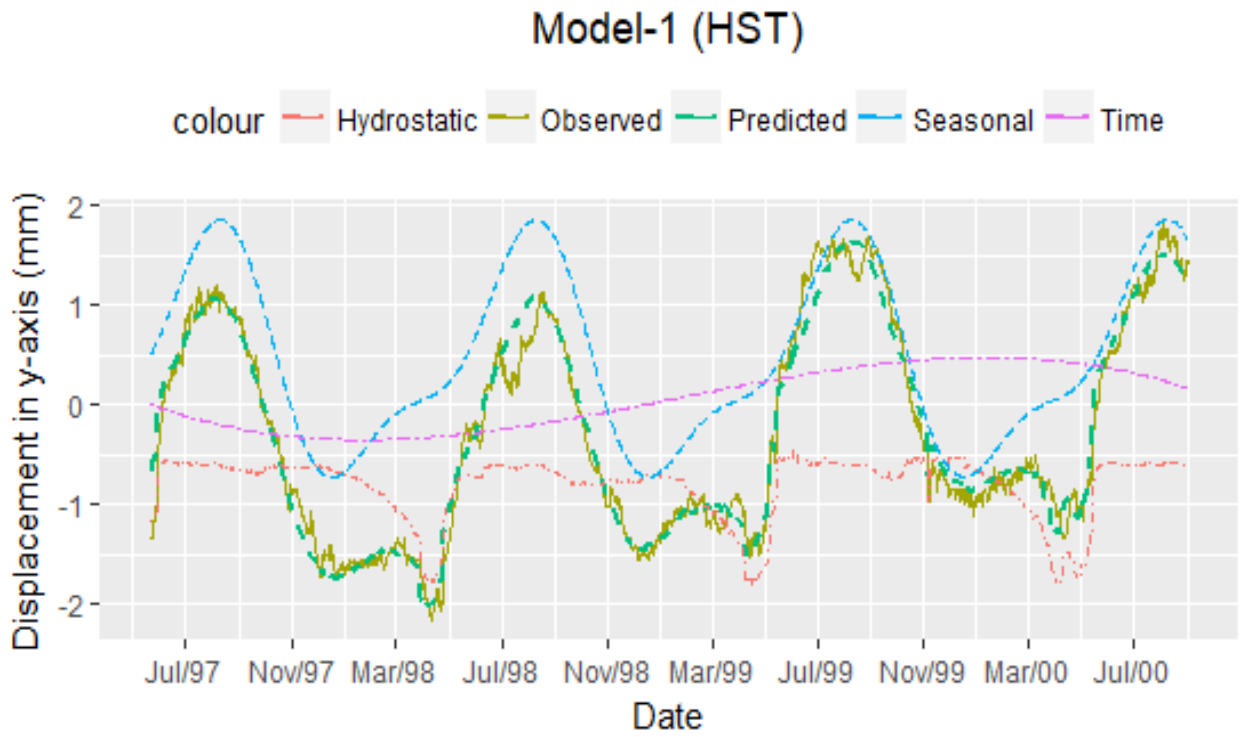

Figure 4.16 - Comparison between the predicted and the observed displacements in the y-axis

In the y-axis or the radial axis, the displacements were more influenced by the hydrostatic term than in the x-axis. Hence the fluctuations caused due to the seasonal component of the model were not highly characteristic in the overall displacement. Therefore, we see a better fit than in the x-axis. However, slight variations were seen during August, which was due to the variability in the daily air temperature.

Figure 4.17 shows that the predicted displacement in the vertical axis is also slightly deviated from the observed displacement during April which is also due to the changes in the daily air temperature unaccounted by the seasonal term of the model.

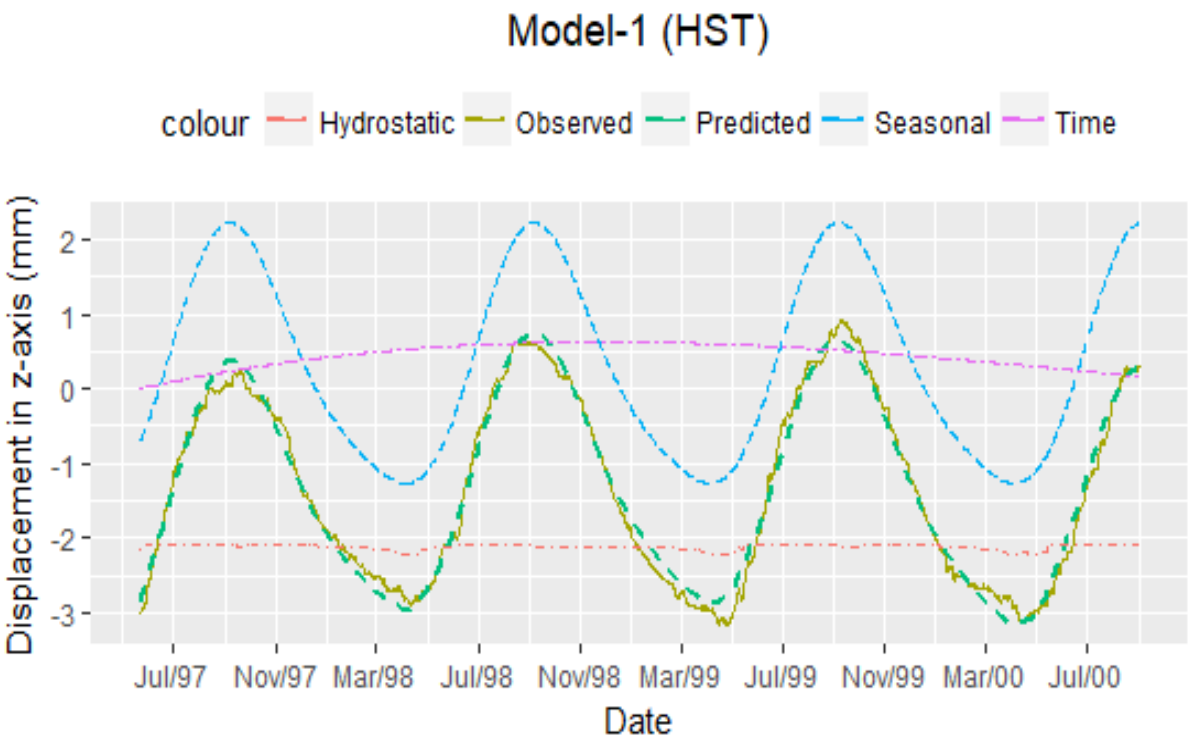

Figure 4.17- Comparison between the predicted and the observed displacements in the z-axis

Residual diagnosis is an integral part of model validation. Skipping residual analysis can lead to spurious regression models creating larger standard errors for the coefficients. In this study, residual plots, histograms and q-q plots of the residuals were plotted for all three axes to check the assumption of normal distribution and white noise.

Figure 4.18, 4.19, & 4.20 show the histogram and the q-q plots for the residuals in the x, y, & z-axes.

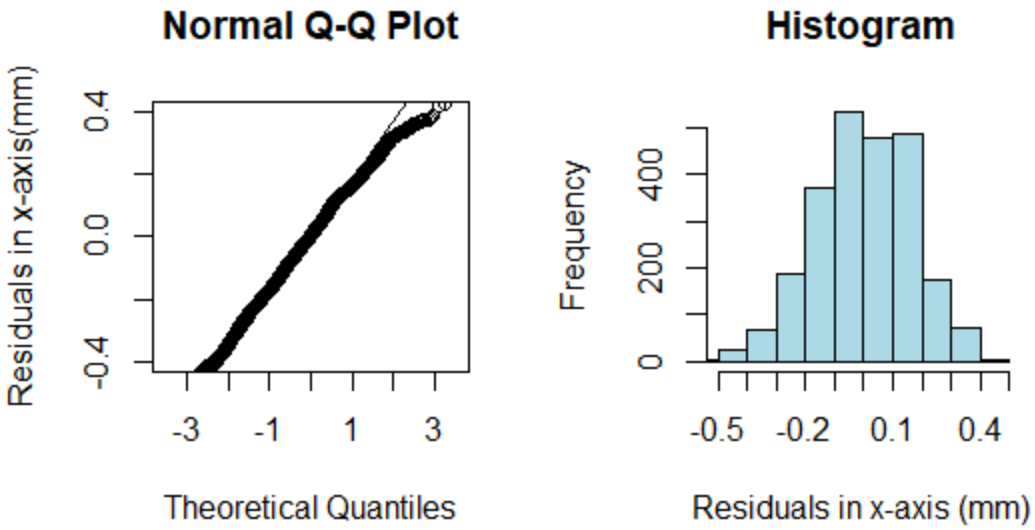

Figure 4.18 - Check for normality of residuals in x-axis

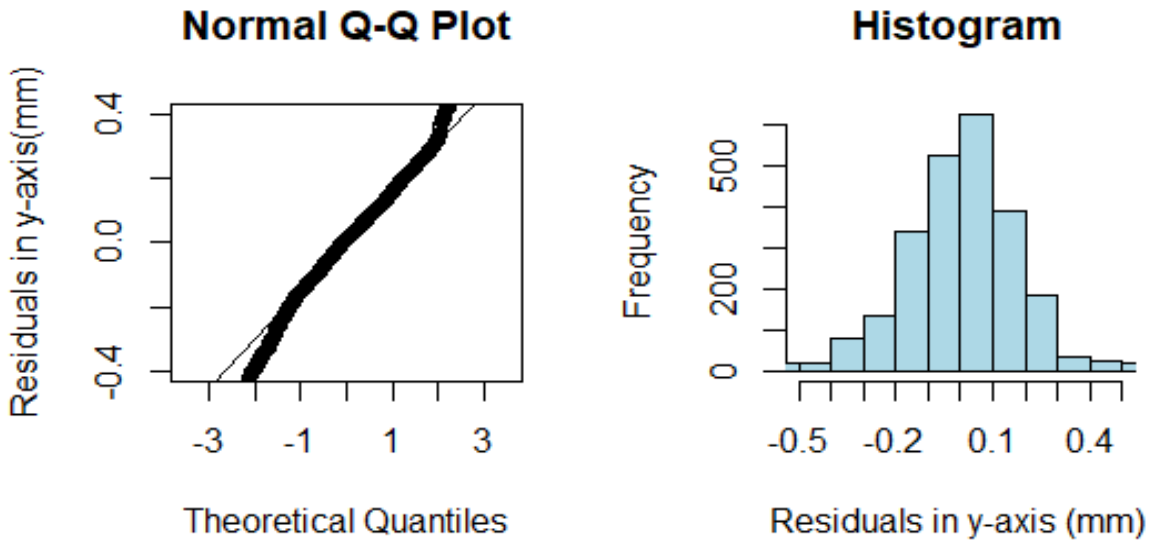

Figure 4.19 - Check for normality of residuals in y-axis

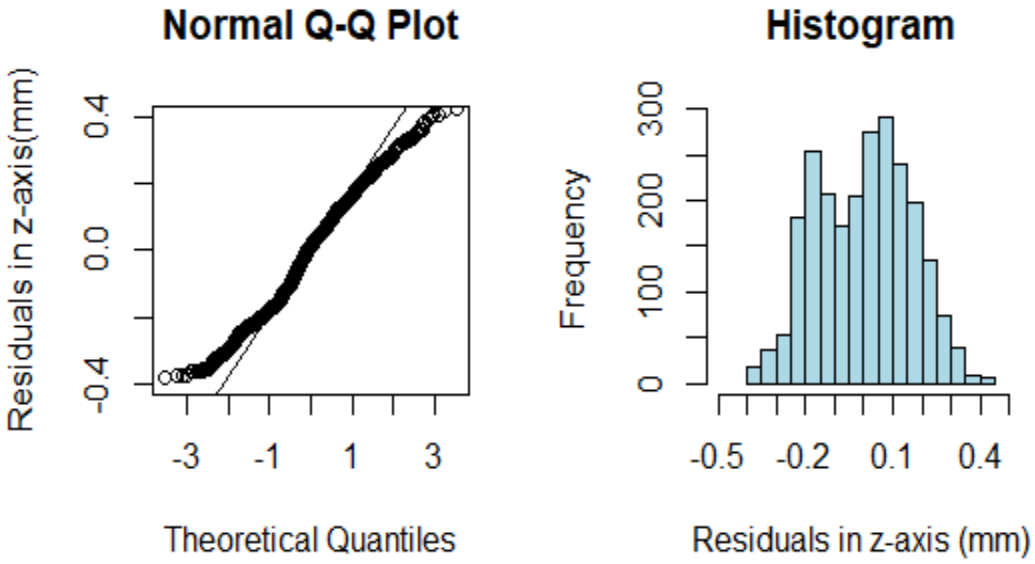

Figure 4.20 - Check for normality of residuals in z-axis

It can be noticed that the residuals for the z-axis are not precisely normally distributed since the histogram is skewed and not all of the residuals in the q-q plot lie along the theoretical normal distribution line. The residuals for the x-axis & the y-axis seem to be normally distributed with a few residuals in the ends not lying on the normal distribution line.

As discussed in 4.2.2.4, the most crucial step of model diagnosis is residual analysis. Therefore, residual plots of the model concerning the fitted displacements have been plotted to confirm the hypothesis of random (white noise) residuals with zero mean and constant variance.

Figures 4.21, 4.22, and 4.23 show the residual plots of the model as a function of the fitted displacements for x, y, and z-axis.

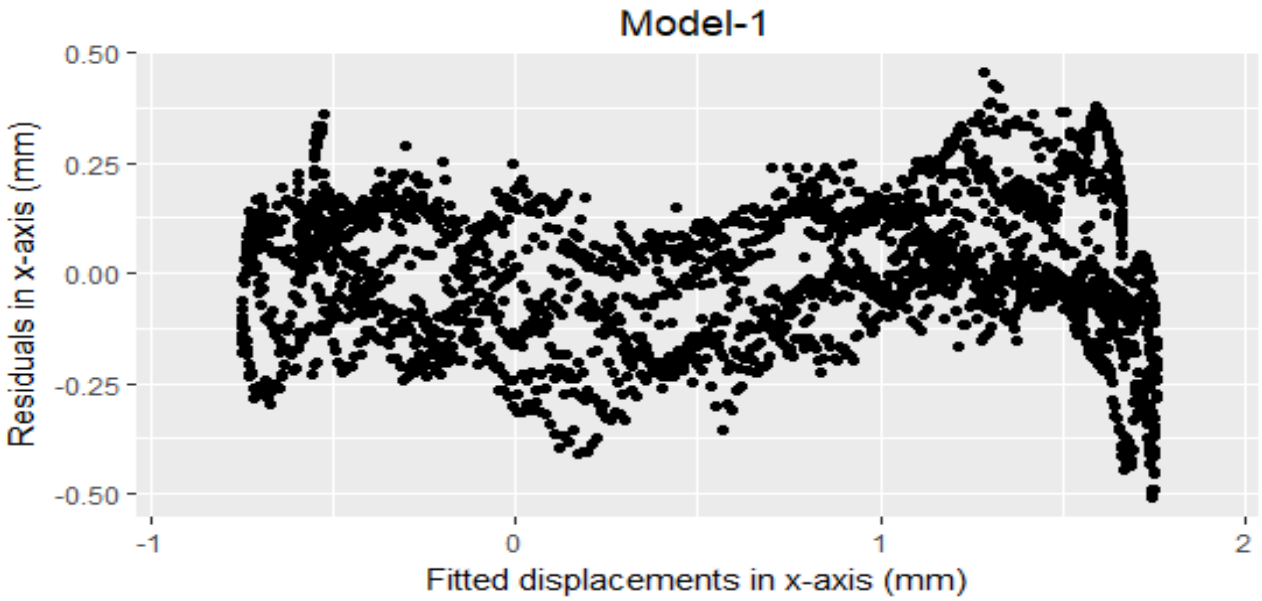

Figure 4.21 - Residual plot of model-1 as a function of fitted displacement in the x-axis

It is evident that the residual plot is not white noise and has a predictive pattern in it. We can see more significant residuals near the right end of the plot, where the fitted displacement is in between 1.5mm to 2mm. This can also be seen in figure 4.15 where the predicted displacements in the x-axis do not satisfactorily fit the observed displacements during March and April.

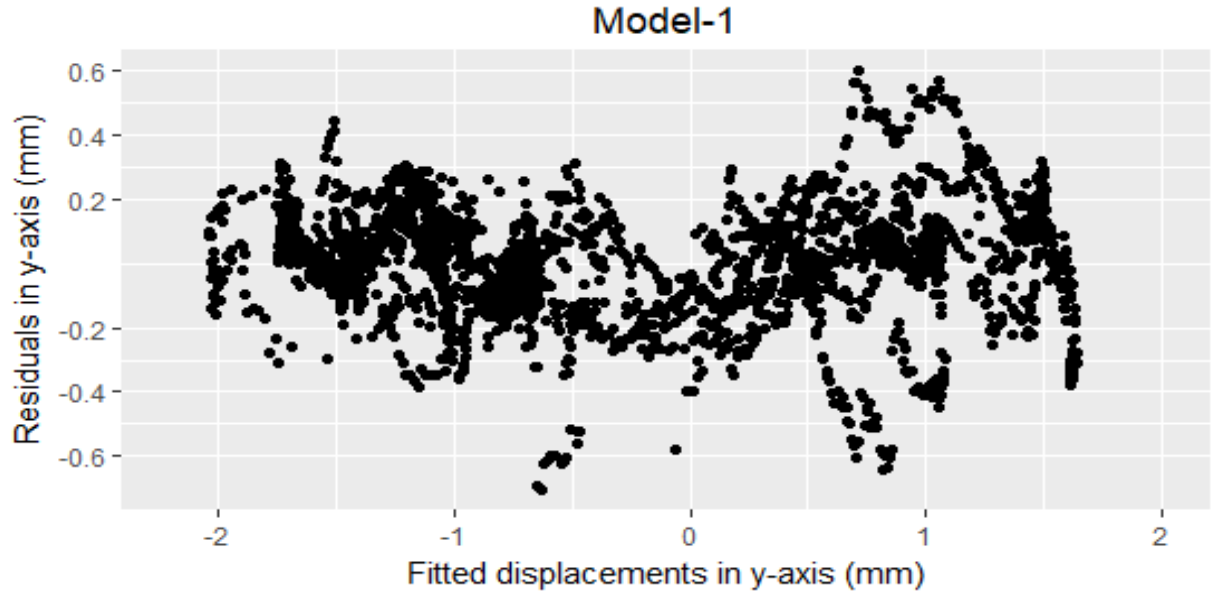

Figure 4.22 - Residual plot of model-1 as a function of fitted displacement in the y-axis

Figure 4.25 depicts that there is a slight trend in the residuals with few large residuals around fitted displacement = 0.8 to 0.9. These larger residuals are due to the imperfect fit for August/98. The mean of the residuals is approximately zero, and the variance is almost constant throughout.

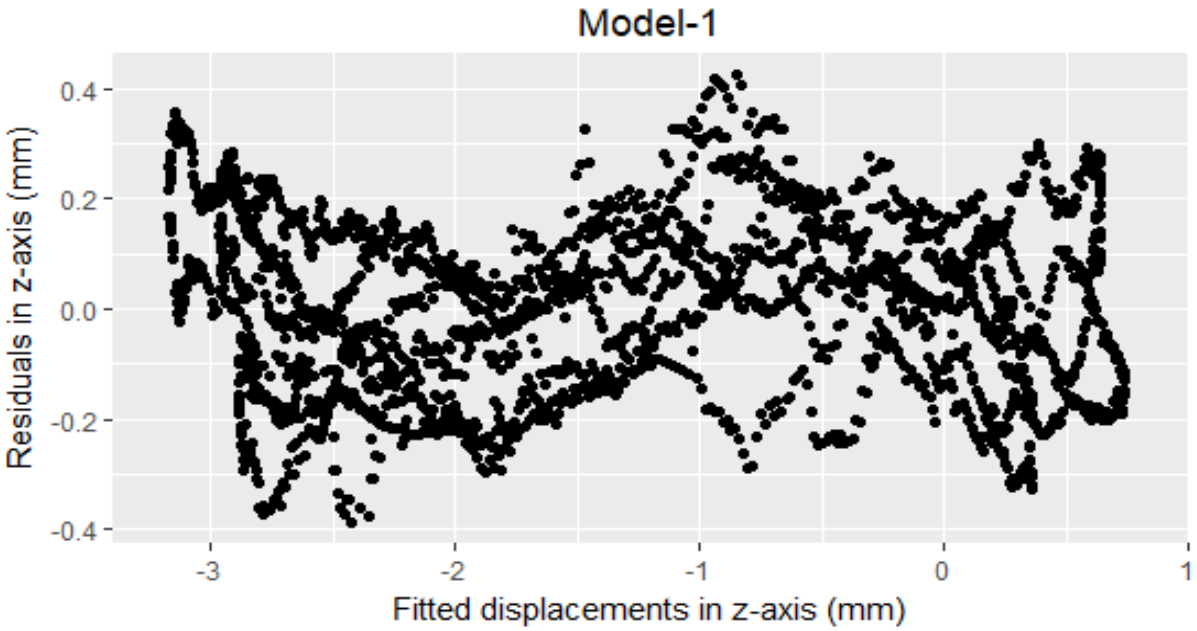

Figure 4.23 - Residual plot of model-1 as a function of fitted displacement in the y-axis

The residual plot in the vertical axis shows a strong trend with a mean of the residuals centered around zero and a constant variance. Therefore we can say that some predictive power is remaining in the residuals. Overall, model – 1 (HST) fits the data accurately except for the days when the air temperature is very high or low as compared with the seasonal component of the model.

4.3.2 Model – 2 (HST Segmented)

As discussed before in section 4.2.2.3, it was found that the relationship between the reservoir level and the displacement is non-linear for all three axes. This non-linear relationship violates the assumption of linearity between the independent and dependent variables and can cause improper fit with larger residuals. We also found that the displacements have increasing amplitudes, and are non-stationary. The standard HST model does not incorporate for non-stationarity due to increasing variance, and neither does it incorporate the effect of the non-linear relationship between the reservoir level and the displacements.

Therefore, two changes have been made to the model-1 to include these effects.

1) To reduce the effect of non-linearity the standardized reservoir level variable (Z) was fitted piecewise to the displacement using a segmented regression. Multi-linear segmented (or piecewise) regression analysis fits the non-linear relationship piecewise, with the help of interaction terms or dummy variables. In this case, the breakpoints of the curve can be approximated by looking at the scatterplot between the dependent and the independent variable. Therefore, the non-linear relationship can be divided into segments of linear relationships. Dummy terms of the variable showing non-linearity are introduced in the model, which change the slope of the regression curve after an estimated breakpoint. These dummy variables only partake in the regression analysis after their breakpoint. This way the non-linear relationship can be modeled better by dividing it into segments of linear relationships.

Figure 4.24 and 4.25 show the difference between the segmented (or piecewise) regression and the standard HST fit.

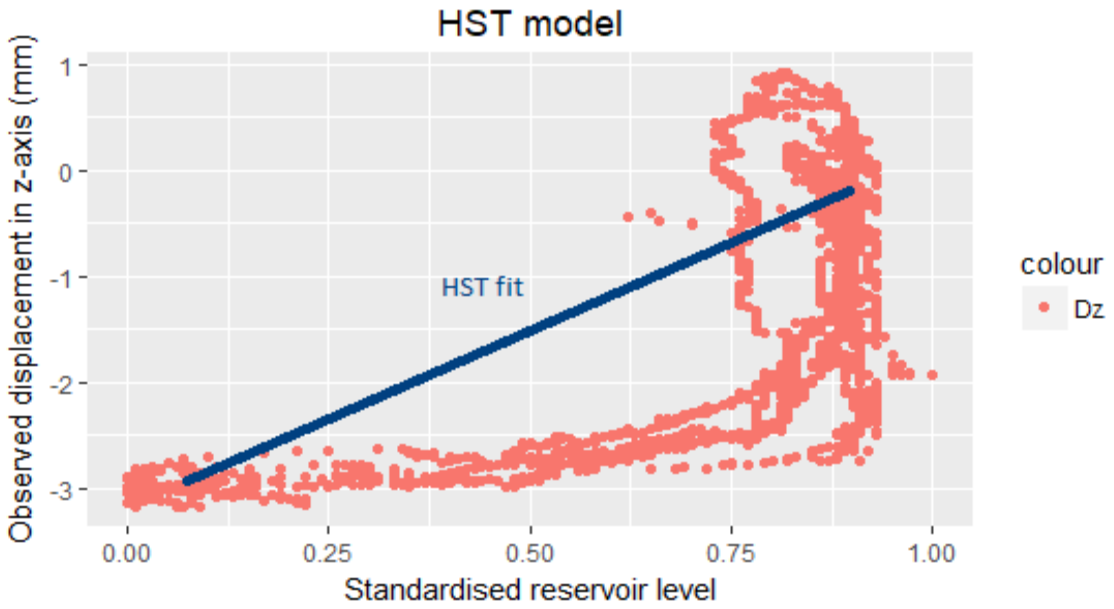

Figure 4.24 - Regular fit for a non-linear relationship

The above figure explains how a standard HST model tries to fit these two variables linearly, whereas the relationship between them is non-linear. This will lead to an improper fit and larger residuals.

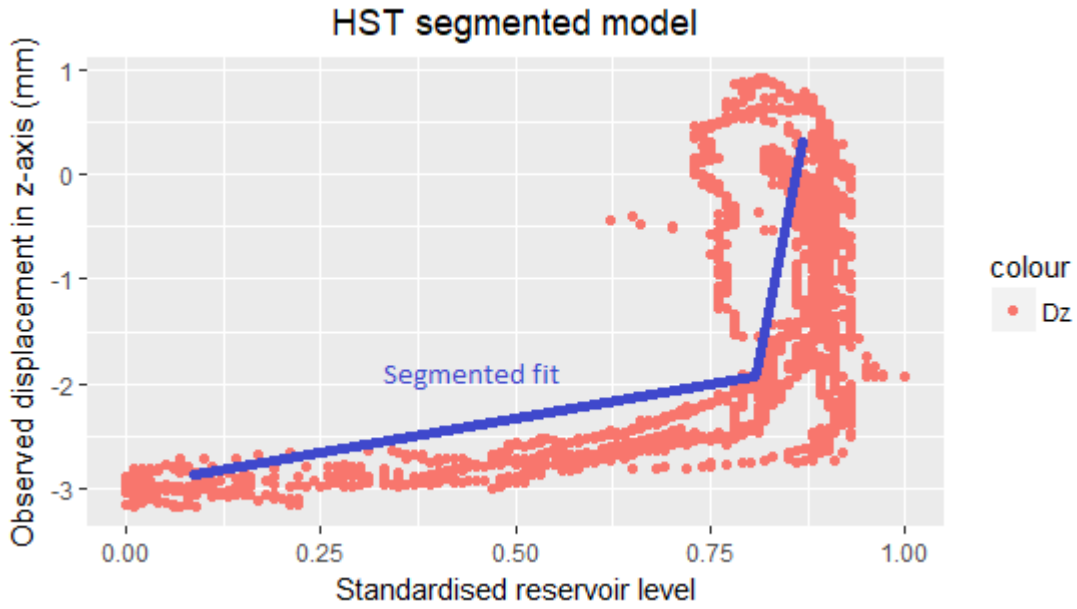

Figure 4.25 - Segmented fit for a non-linear relationship

It is clear from the Figure 4.25 that a piecewise fit provides a better fit for the hydrostatic part of the model as compared to model-1.

2) A linear term of time was multiplied by the seasonal component of the model which can incorporate for the increasing amplitudes of the displacement curve.

The previous seasonal term was: $a_2*\sin(T) + a_3*\cos(T) + a_4*\sin(T)*\cos(T) + a_5*\sin^2(T)$

The new proposed seasonal term is: $(1+t)*(a_2*\sin(T) + a_3*\cos(T) + a_4*\sin(T)*\cos(T) + a_5*\sin^2(T))$

⇨ $a_2*\sin(T) + a_3*\cos(T) + a_4*\sin(T)*\cos(T) + a_5*\sin^2(T) + t*(a_2*\sin(T) + a_3*\cos(T) + a_4*\sin(T)*\cos(T) + a_5*\sin^2(T))$

The second part of the seasonal term with interaction terms between time and the trigonometrical functions is assumed to capture the increment in the amplitudes of the displacement.

After this transformation the model looks like:

HSTseg_Z → $Dz \sim a_0 + a_1* Z_i + a_2*\sin(T) + a_3*\cos(T) + a_4*\sin(T)*\cos(T) + a_5*\sin^2(T) + t*a_6*\sin(T) + t*a_7*\cos(T) + t*a_8*\sin(T)*\cos(T) + t*a_9*\sin^2(T) + a_{10}*t + a_{11}*t^2 + a_{12}*t^3$ 4.16

HSTseg_Y → $Dy \sim a_0 + a_1* Z_i + a_2*\sin(T) + a_3*\cos(T) + a_4*\sin(T)*\cos(T) + a_5*\sin^2(T) + t*a_6*\sin(T) + t*a_7*\cos(T) + t*a_8*\sin(T)*\cos(T) + t*a_9*\sin^2(T) + a_{10}*t + a_{11}*t^2 + a_{12}*t^3$ 4.17

HSTseg_X → $Dx \sim a_0 + a_1* Z_i + a_2*\sin(T) + a_3*\cos(T) + a_4*\sin(T)*\cos(T) + a_5*\sin^2(T) + t*a_6*\sin(T) + t*a_7*\cos(T) + t*a_8*\sin(T)*\cos(T) + t*a_9*\sin^2(T) + a_{10}*t + a_{11}*t^2 + a_{12}*t^3$ 4.18

Stepwise regression was used to calculate the coefficients of the regression equations. Only the hydrostatic component undergoes piecewise fit and the rest of the components are fitted linearly

using a normal stepwise algorithm. Tables 4.4, 4.5, and 4.6 depict the regression summaries of the segmented regression in radial, tangential, and vertical axis respectively.

Independent variables	Coeff.Estimate	Std.Error	t-value	p-value(>|t|)
Intercept	1.263e+00	2.150e-02	58.755	< 2e-16
sin (T)	-1.004e+00	1.149e-02	-87.354	< 2e-16
cos (T)	2.448e-01	1.010e-02	24.243	< 2e-16
sin (T).cos(T)	-5.250e-02	1.978e-02	-2.654	0.008
$\sin^2$ (T)	-4.899e-01	1.810e-02	-27.070	< 2e-16
t.sin (T)	-8.410e-05	1.608e-05	-5.230	1.85e-07
t.cos (T)	1.758e-04	1.546e-05	11.366	< 2e-16
t.sin (T).cos (T)	1.375e-04	2.818e-05	4.882	1.12e-06
t.$\sin^2$ (T)	2.871e-05	2.654e-05	1.082	0.279
t	-1.584e-03	1.261e-04	-12.569	< 2e-16
t^2	2.079e-06	2.415e-07	8.610	< 2e-16
t^3	-7.862e-10	1.314e-10	-5.982	2.53e-09
Residual standard error			0.1516 on 2380 degrees of freedom	
Multiple R-squared			0.9681	
Adjusted R-squared			0.9679	
Predicted R-squared			0.9677	

Table 4.4 - Regression summary for model-2 in tangential direction (x-axis)

The regression summary for x-axis shows that the stepwise algorithm for model-2 did not include the standard reservoir level in the final model. Therefore a segmented fit of the reservoir level does not create any difference in the results for the x-axis. Since the t-values for the seasonal terms are high, we can say that they dominate the response in the x-axis. The adjusted R-squared was found to increase slightly from 0.9619 to 0.9679, and the residual standard error reduced from 0.165 to 0.1516 (~8% reduction), which was due to the transformation made in the seasonal component of the model.

Independent variables	Coeff.Estimate	Std.Error	t-value	p-value(>|t|)
Intercept	-1.680e+00	2.979e-02	-56.403	< 2e-16
Z	1.042e+00	5.085e-02	20.494	< 2e-16
Sin (T)	8.498e-01	1.337e-02	63.566	< 2e-16
Cos (T)	3.564e-01	1.099e-02	32.440	< 2e-16
Sin (T).Cos(T)	-4.153e-02	2.228e-02	-1.865	0.0624
Sin^2 (T)	7.519e-01	1.919e-02	39.173	< 2e-16
t.sin (T)	1.085e-04	1.667e-05	6.510	9.13e-11
t.cos (T)	3.029e-04	1.633e-05	18.544	< 2e-16
t.sin (T).cos (T)	-7.480e-05	3.005e-05	-2.489	0.0129
$t.sin^2$ (T)	2.577e-04	2.791e-05	9.232	< 2e-16
t	-4.160e-03	1.272e-04	-32.694	< 2e-16
t^2	9.927e-06	2.441e-07	40.666	< 2e-16
t^3	-5.560e-09	1.339e-10	-41.518	< 2e-16
U1.Z	9.697e-01	8.525e-02	11.375	NA
Estimated Break-Point in Z			0.511	
Residual standard error			0.1598 on 2380 degrees of freedom	
Multiple R-squared			0.9773	
Adjusted R-squared			0.9771	
Predicted R-squared			0.9769694	

Table 4.5 - Regression summary for model-2 in radial direction (y-axis)

The regression summary above shows a new variable U1.Z, which is the dummy variable used to perform piecewise regression analysis on the standard reservoir level. Its coefficient denotes the difference between the slopes of the two piecewise linear fits made for the standardized reservoir level. The R-squared of the model-2 was slightly better than model-1, and the residual standard error dropped from 0.177 to 0.1598 (~10% reduction). Therefore it can be concluded that model-2 is better at reducing the residual standard error. The estimated breakpoint for the piecewise fit converged at Z = 0.507. Figure - 4.9 shows that the breakpoint is around reservoir level = 100m, which can be converted to standardized reservoir level of 0.75 using the formula:

$$Z = \frac{\mathrm{Ht} - \mathrm{Hmin}}{\mathrm{Hmax} - \mathrm{Hmin}}$$

It was not possible to further improve the estimate of the breakpoint in y-axis since the segmented function did not converge for Z=0.75

Independent variables	Coeff.Estimate	Std.Error	t-value	p-value(>|t|)
Intercept	-2.256e+00	2.625e-02	-85.937	< 2e-16
Z	9.033e-02	2.639e-02	3.423	0.000631
Sin (T)	1.462e+00	1.119e-02	130.586	< 2e-16
Cos (T)	-6.828e-01	9.772e-03	-69.875	< 2e-16
Sin (T).cos(T)	1.640e-01	1.878e-02	8.733	< 2e-16
Sin^2 (T)	4.788e-01	1.689e-02	28.345	< 2e-16
t.sin (T)	9.616e-05	1.510e-05	6.366	2.32e-10
t.cos (T)	-6.805e-05	1.453e-05	-4.683	2.99e-06
t.sin (T).cos (T)	-4.825e-04	2.664e-05	-18.111	< 2e-16
$t.sin^2$ (T)	2.029e-04	2.486e-05	8.162	5.30e-16
t	3.000e-03	1.177e-04	25.488	< 2e-16
t^2	-4.023e-06	2.261e-07	-17.789	< 2e-16
t^3	1.349e-09	1.231e-10	10.958	< 2e-16
U1.Z	1.214e+00	1.312e-01	9.251	NA
Estimated breakpoint			0.796	
Residual standard error			0.1413 on 2380 degrees of freedom	
Multiple R-squared			0.9869	
Adjusted R-squared			0.9868	
Predicted R-squared			0.9867	

Table 4.6 - Regression summary for model-2 in vertical direction (z-axis)

The estimated breakpoint for model-2 in the vertical axis was 0.796, and from the figure 4.25, it can be seen that the breakpoint is near Z = 0.8. The R-squared for model-2 increased slightly and the standard residual error reduced from 0.1597 to 0.1413 (~11.5% reduction). Since the analysis is for the vertical axis, seasonal terms dominate again with t-values >100 for sin(T), whereas the standard reservoir level has a t-value of only 3.423. Therefore, the segmented fit of the reservoir level does not produce much effect on the model accuracy. Most of the reduction in error is due to the interaction terms of time and seasonal components, which incorporate for the increasing amplitudes of the displacement curve.

Further, figures 4.26, 4.27, and 4.28 compare the predicted displacements from model-2 with the observed displacements for x, y, and z-axes respectively. Further, the displacements have been decomposed into three components of the model to better visualize the effects of these components on the global displacement.

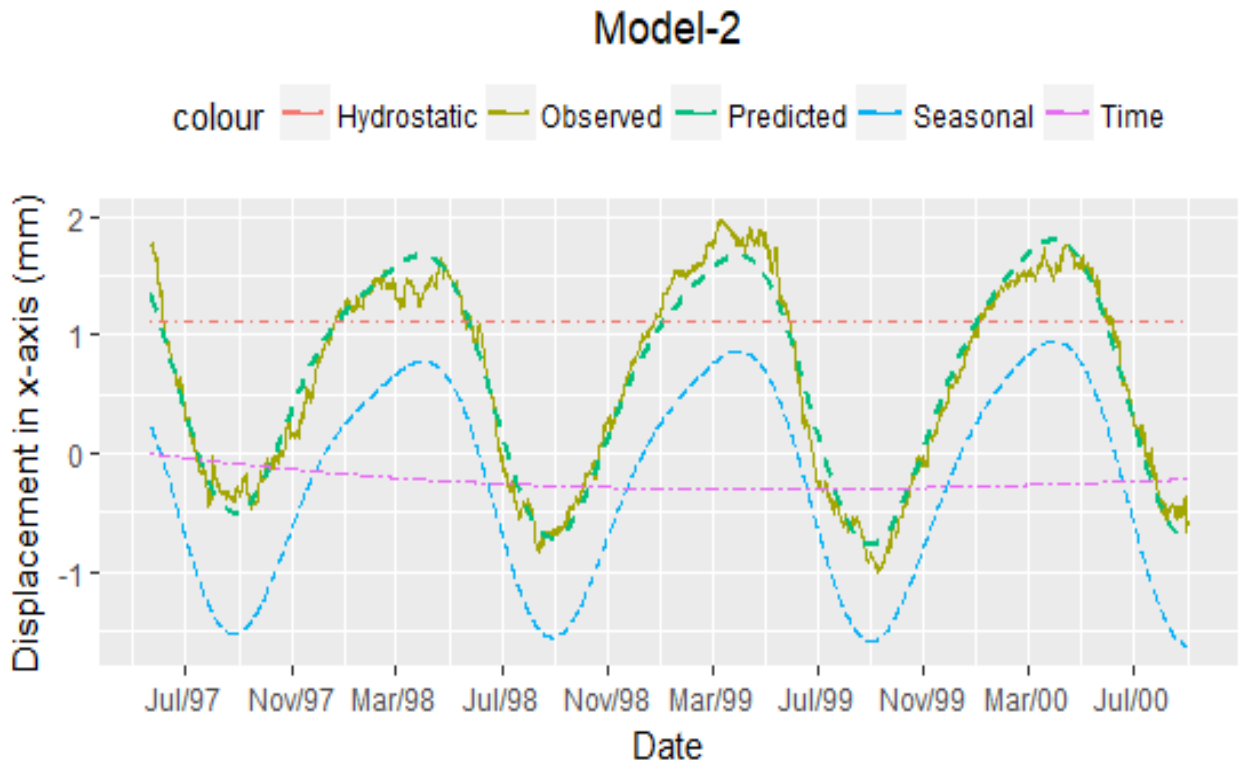

Figure 4.26 - Comparison between predicted and observed displacements for model-2 in x-axis

The above figure shows that the amplitude of the predicted curve keeps on increasing with time, which slightly improves the fit for March and September.

It can also be noticed that the hydrostatic curve is constant since it was not included in the model and only depicts the intercept of the model. The seasonal curve denoted by the blue curve is the most influencing factor in modeling the displacement in the x-axis.

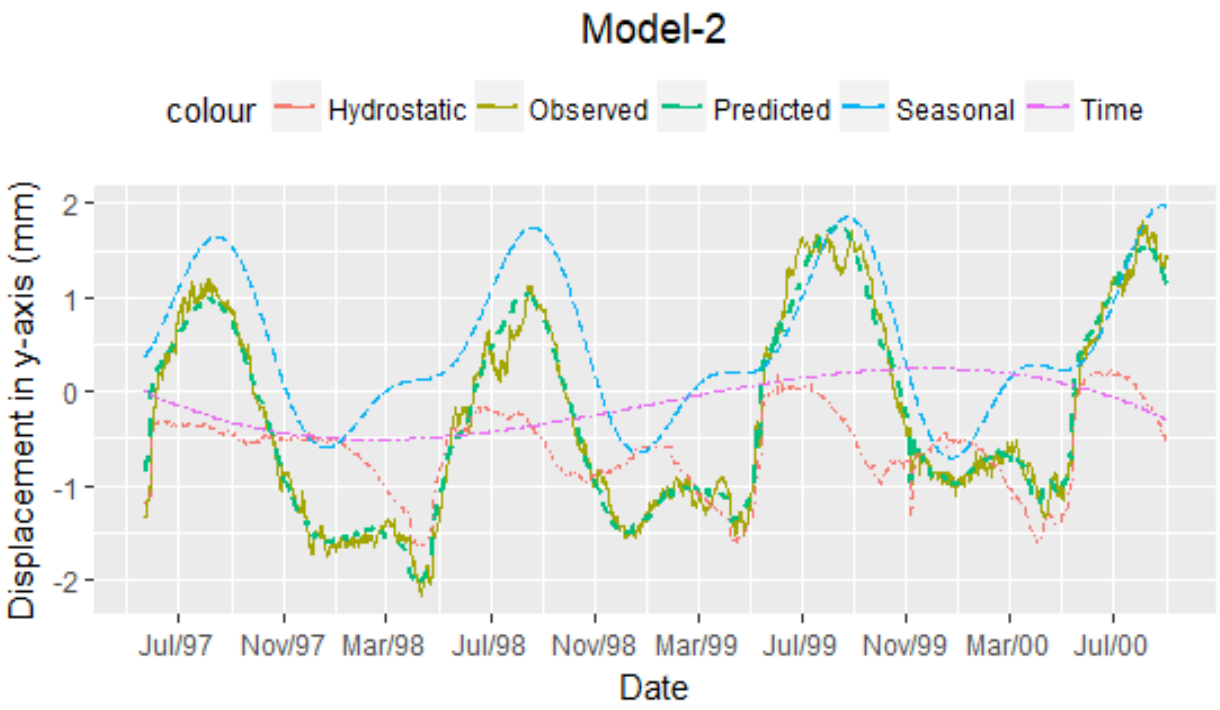

Figure 4.27 - Comparison between predicted and observed displacements for model-2 in y-axis

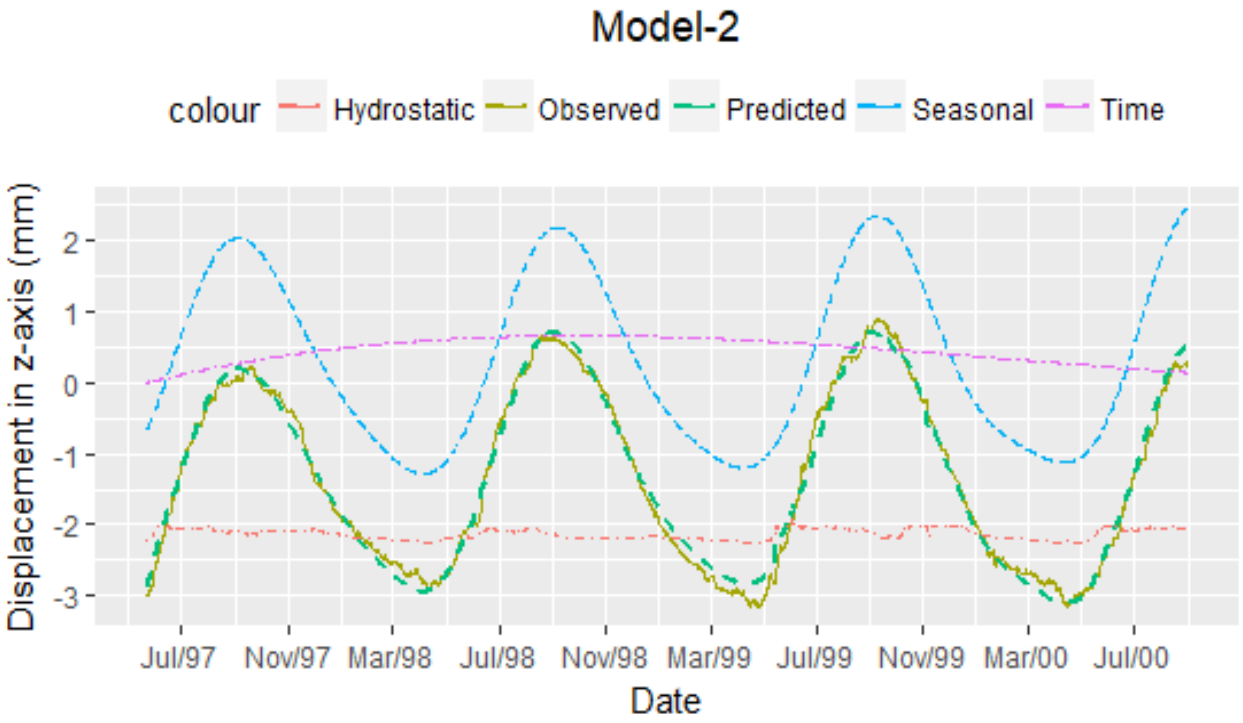

Figure 4.28 - Comparison between predicted and observed displacements for model-2 in z-axis

Figure-4.28 shows that model-2 slightly improves the fit in z-axis than model-1 for September during all three years.

To show the effects of a piecewise fit of the standardized reservoir level plots between the displacements due to hydrostatic component and the standardized reservoir levels have been made for y and z-axis. Figure 4.29 & 4.30 depicts the plot of the effect of the standardized reservoir level (Z) on the overall displacement and the standardized reservoir level (Z). Segmented regression creates breakpoints of 0.511 for the y-axis and 0.796 for the z-axis.

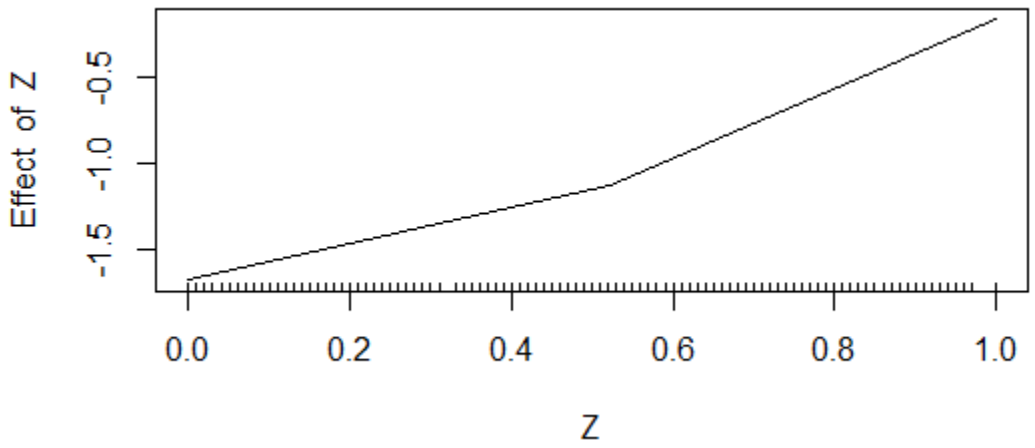

Figure 4.29 - Displacement effect of Z due to segmented regression in y-axis

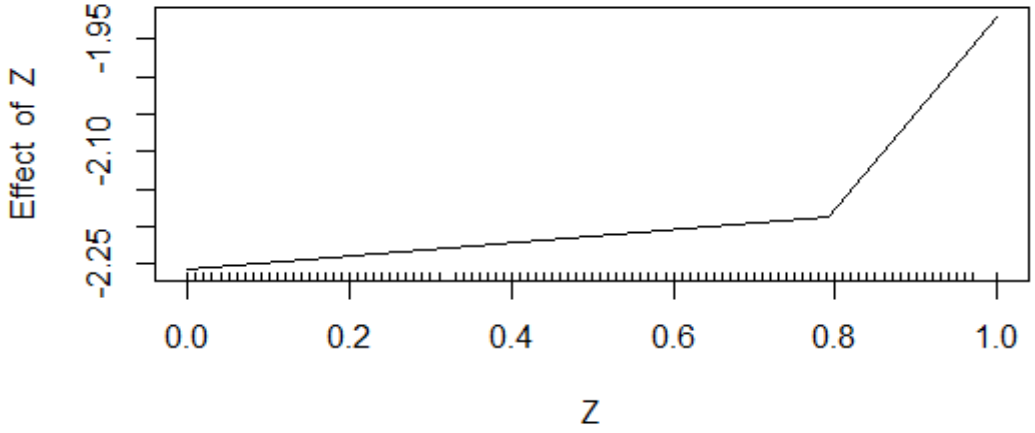

Figure 4.30 - Displacement effect of Z due to segmented regression in z-axis

To check for the normality of the residuals, histograms and q-q plots have been plotted for all three axes.

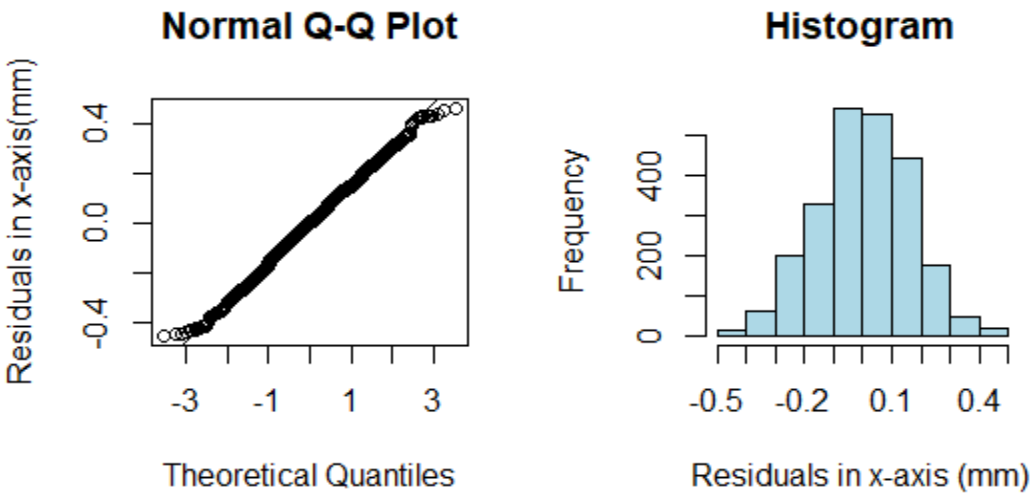

Figure 4.31 - Normality check of residuals for model-2 in x-axis

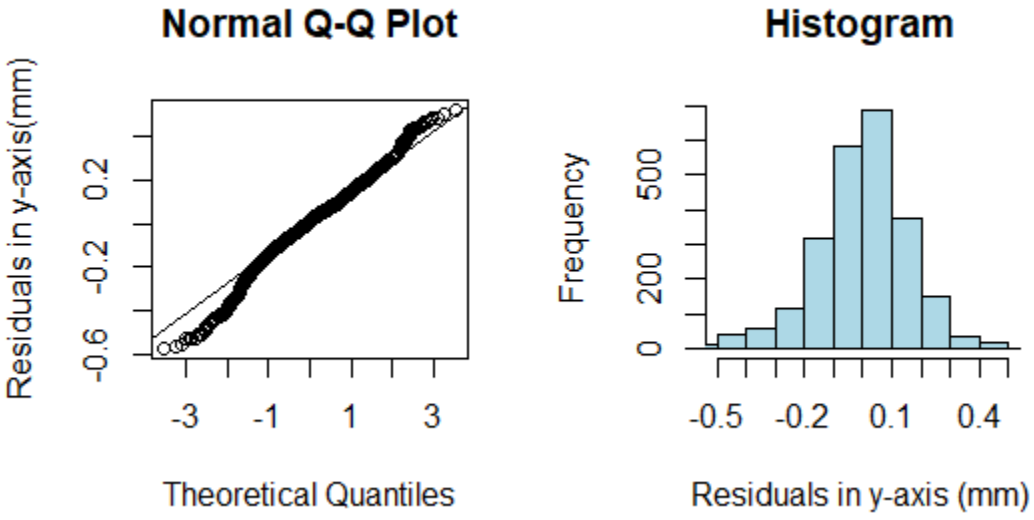

Figure 4.32 - Normality check of residuals for model-2 in y-axis

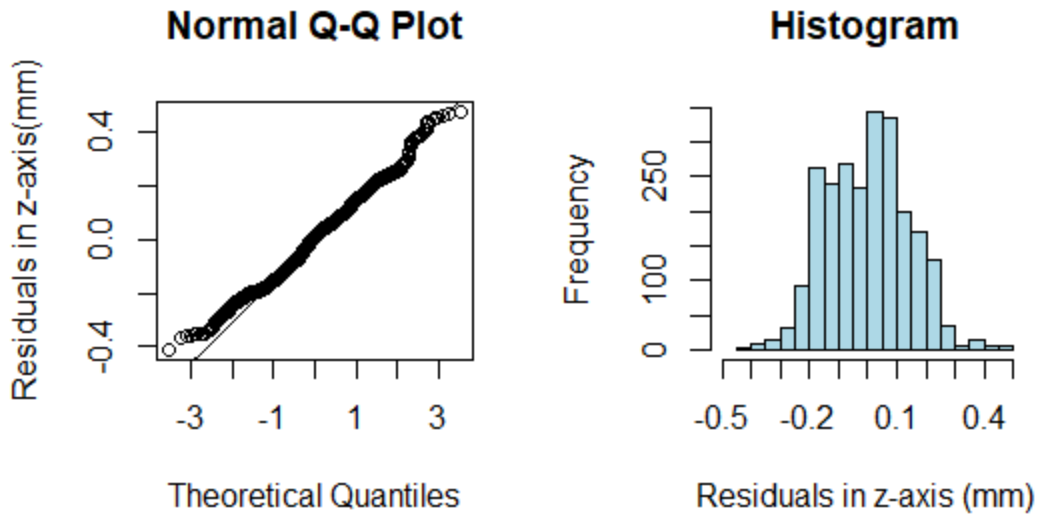

Figure 4.33 - Normality check for residuals of model-2 in z-axis

The residuals in all three axes follow a normal distribution. Model-2 has shown to improve the normality of the residuals in all three axes, especially in the z-axis where the residuals did not follow normal distribution in model-1. Further, the residual analysis was performed to check whether the residuals were white noise or not.

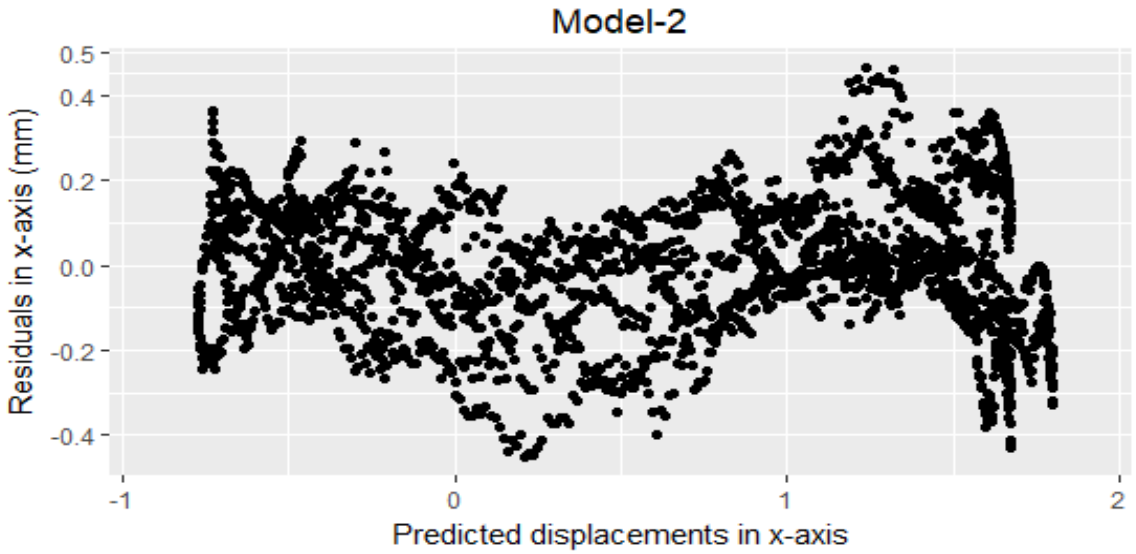

Figure 4.34 - Residuals as a function of the fitted displacement in x-axis for model-2

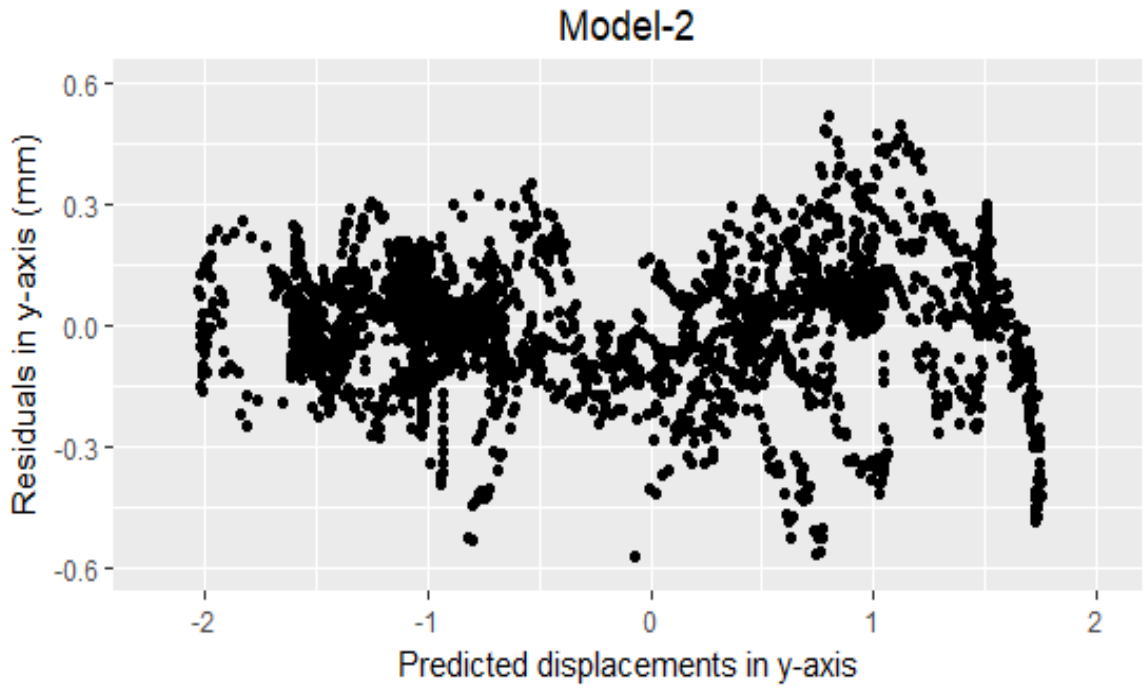

Figure 4.35 - Residuals as a function of the fitted displacement in y-axis for model-2

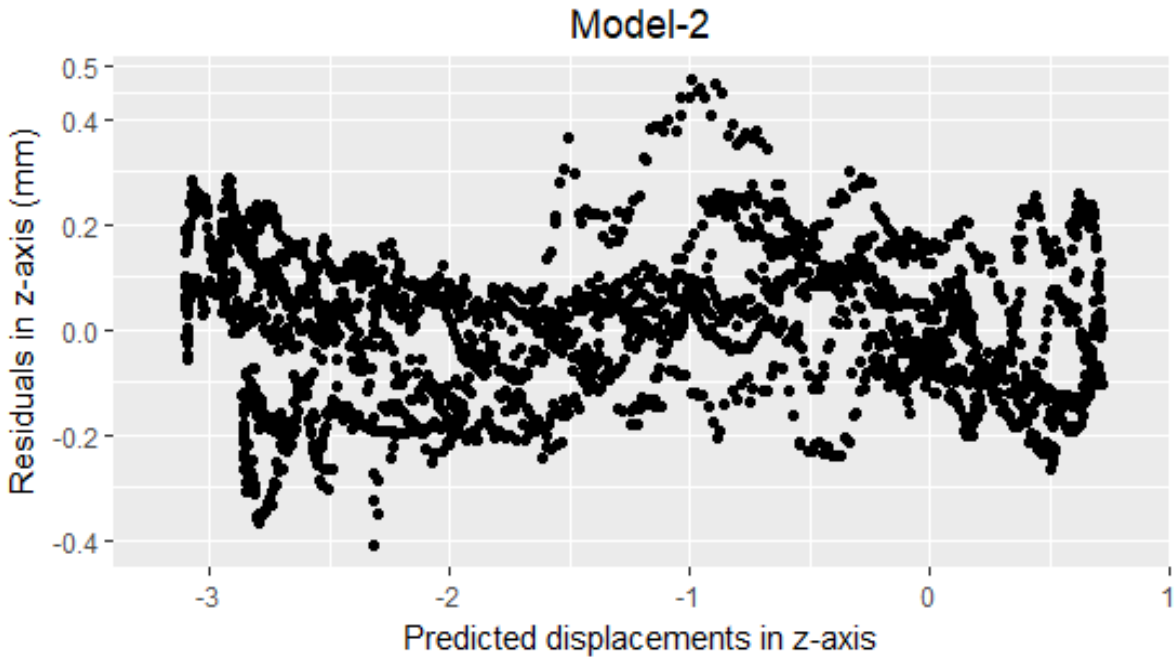

Figure 4.36 - Residuals as a function of the fitted displacement in z-axis for model-2

From the above residual plots, it can be concluded that model-2 does not reduce the trend in the residual curve. However, it reduces the overall residual values for all three axes by (~10%) which is similar to the results of the regression summary where the where the model-2 showed smaller standard error than model-1.

4.3.3 Model-3 (HST_nonparametric)

The previous two models show that the seasonal component of the model does not capture the daily air temperature fluctuations, which leads to larger residuals. Therefore, a non-parametric fit for the seasonal component of the model has been made using locally weighted linear regression (LOESS) to fit these fluctuations better. The non-parametric fit will serve as an empirical seasonal curve which can be added to the HT model to get the predicted displacements.

To capture the seasonal component of the observed data, firstly an HST model was fitted to the data, following which the displacement effects of the seasonal component were separated from the HST model for the entire times series. The remaining model is called the HT model. The residuals of the HT model were calculated which represent the seasonal component and the error. The residuals of the HT model were converted into a yearly seasonal scale (730 observations) by overlapping the residuals for all the years in just one. Further, the empirical seasonal curve was produced using locally weighted linear regression. This empirical seasonal curve will serve the purpose of the trigonometric functions used in the HST model. The displacement effects of the seasonal component can be computed from this curve and can be added to the hydrostatic and time components of the HT model, to predict the overall displacement of the dam in all three axes. The following schematic equations help to explain the procedure:

HST = Hydrostatic component + Seasonal component + Time

↓

HT = HST – Seasonal component

↓

Residuals_HT = Observed displacement – HT = Seasonal component +error

↓

Non-parametric seasonal curve = Loess (seasonal component + error)

↓

Predicted displacements = HT + non-parametric seasonal curve

Figure 4.37, 4.38, & 4.39 show the yearly seasonal curves produced using locally weighted linear regression for x, y, and z-axis respectively.

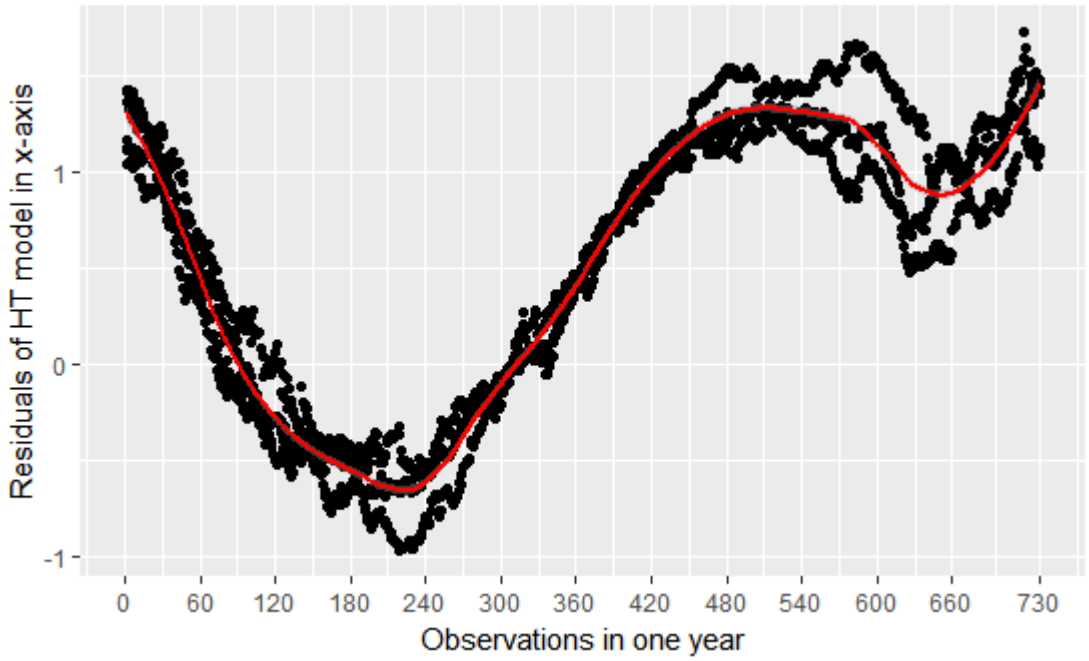

Figure 4.37 - Seasonal curve for x-axis using LOESS fit

Figure 4.38 - Seasonal curve for y-axis using LOESS fit

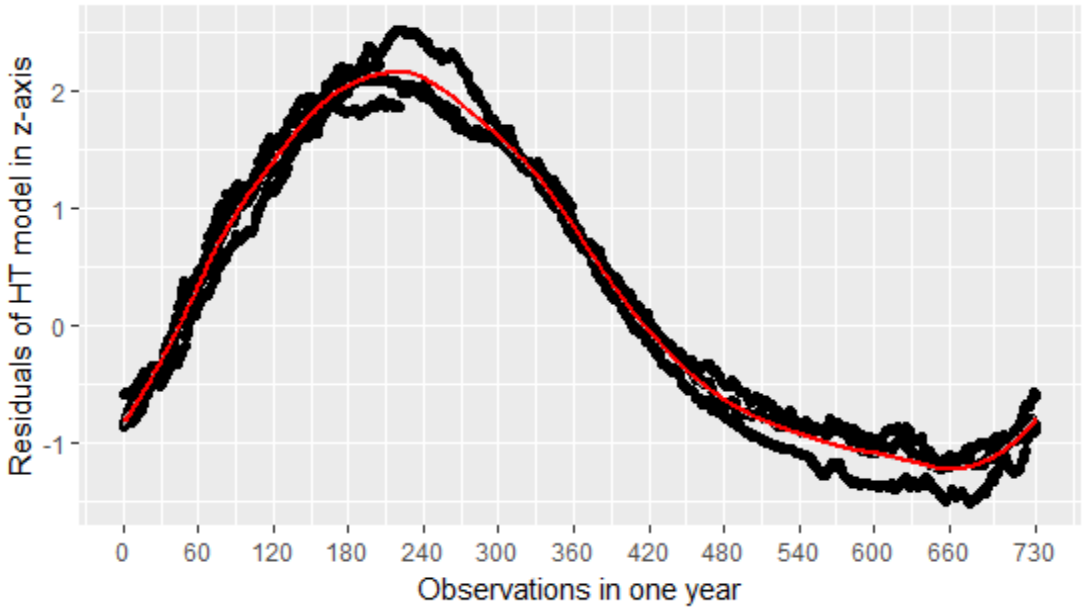

Figure 4.39 - Seasonal curve for z-axis using LOESS fit

The above plots approximate the seasonal component over the years by fitting one curve to it using LOESS fit. This method will help to average out the fluctuations caused due to temperature

variations in any particular year(s) Since LOESS is a non-parametric method, we do not get any parameter estimates but directly the values for the seasonal component of the displacement. This seasonal component can then be added to the hydrostatic and time component of the HT model to predict the overall displacements.

Figure 4.40, 4.41, and 4.42 compares the fit of the seasonal component of the standard HST, and the non-parametric seasonal curve to the residuals of the HT model which describe the seasonal effects for x, y, and z-axis. It was found that the seasonal component of the HST cannot account for the fluctuations caused due to air temperature, whereas the loess fit seasonal curve can better approximate the fluctuations. This improvement is because loess is a non-parametric fit and is more flexible than the seasonal component of the HST model, which is based on smooth trigonometric functions. Lower residuals are expected from the non-parametric model since it accounts for the fluctuations in daily temperature better than the regular seasonal model.

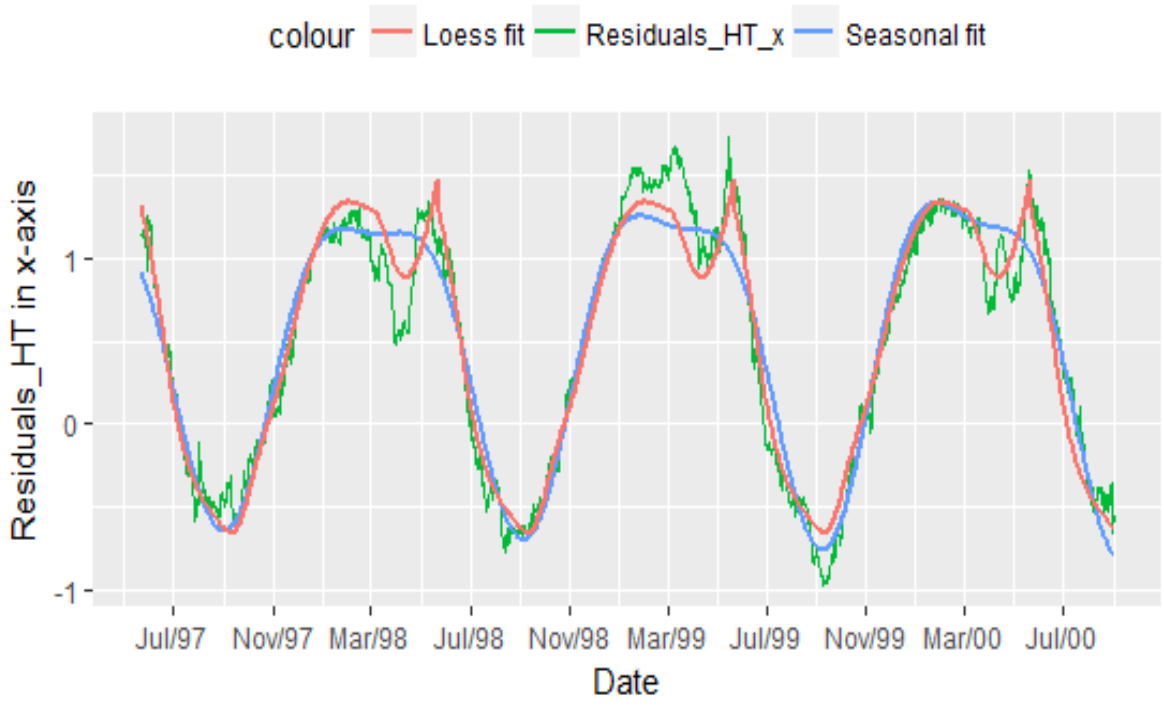

Figure 4.40 - Comparison between different seasonal fits in x-axis

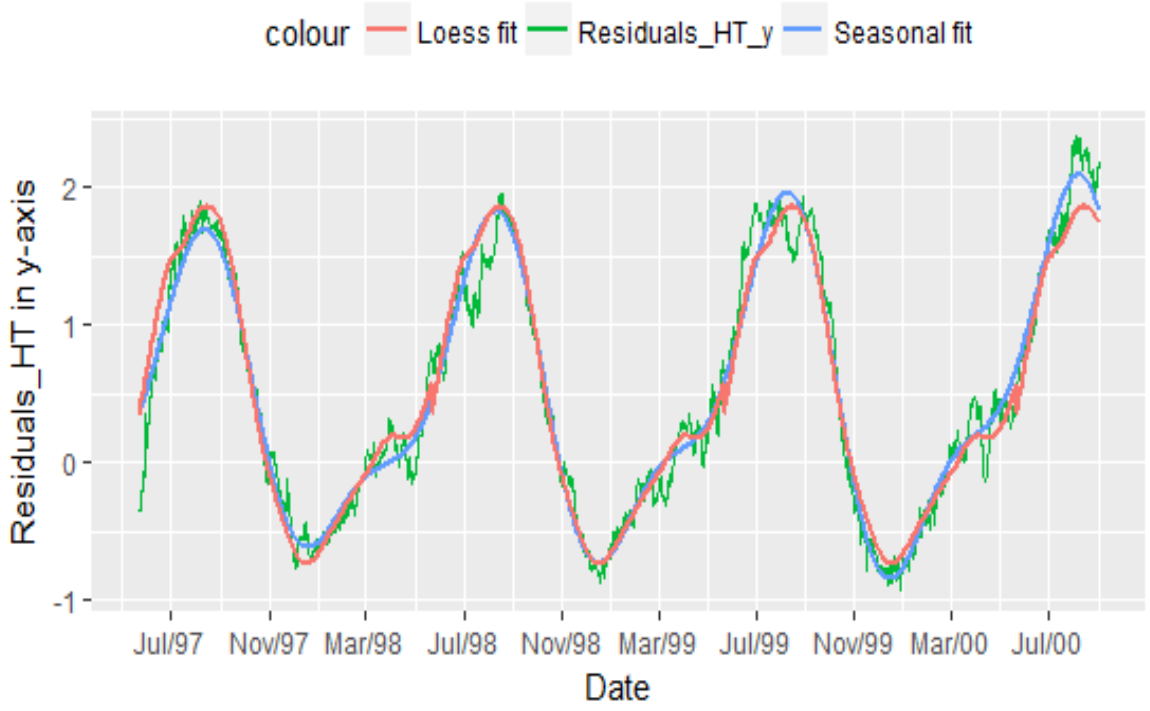

Figure 4.41 - Comparison between different seasonal fits in y-axis

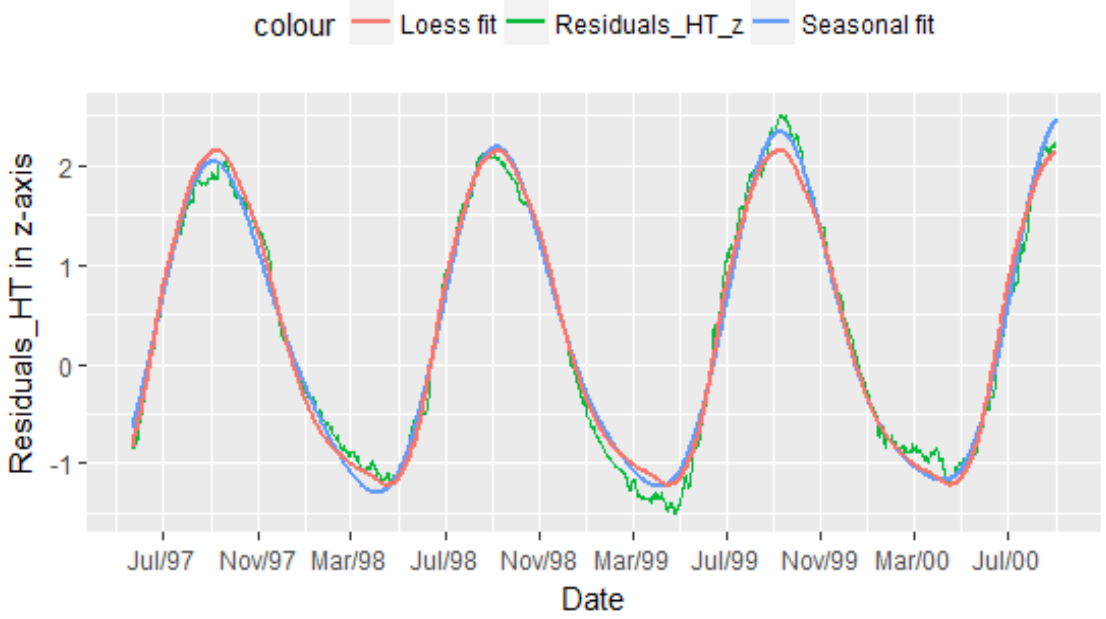

Figure 4.42 - Comparison between different seasonal fits in z-axis

Figure 4.43, 4.44, and 4.45 compare the observed displacements with the predicted displacements obtained from model-3 in x, y, and z-axis.

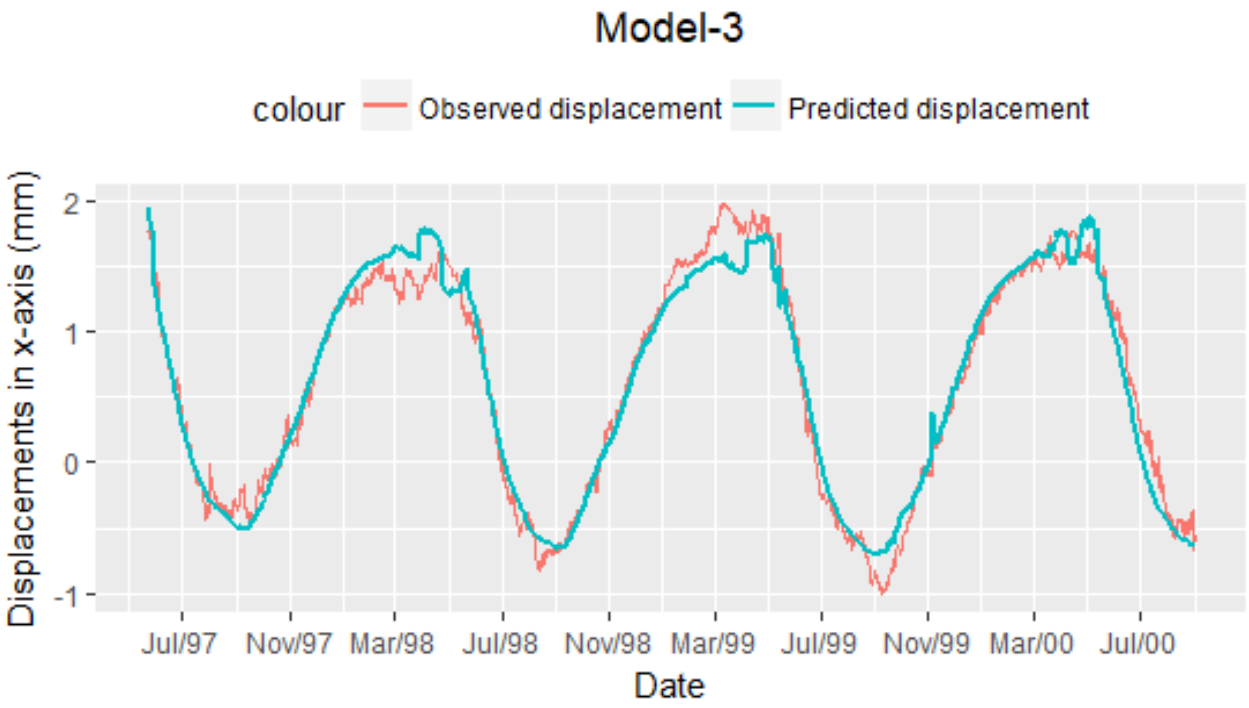

Figure 4.43 -Comparison between observed and predicted displacements for model-3 in x-axis

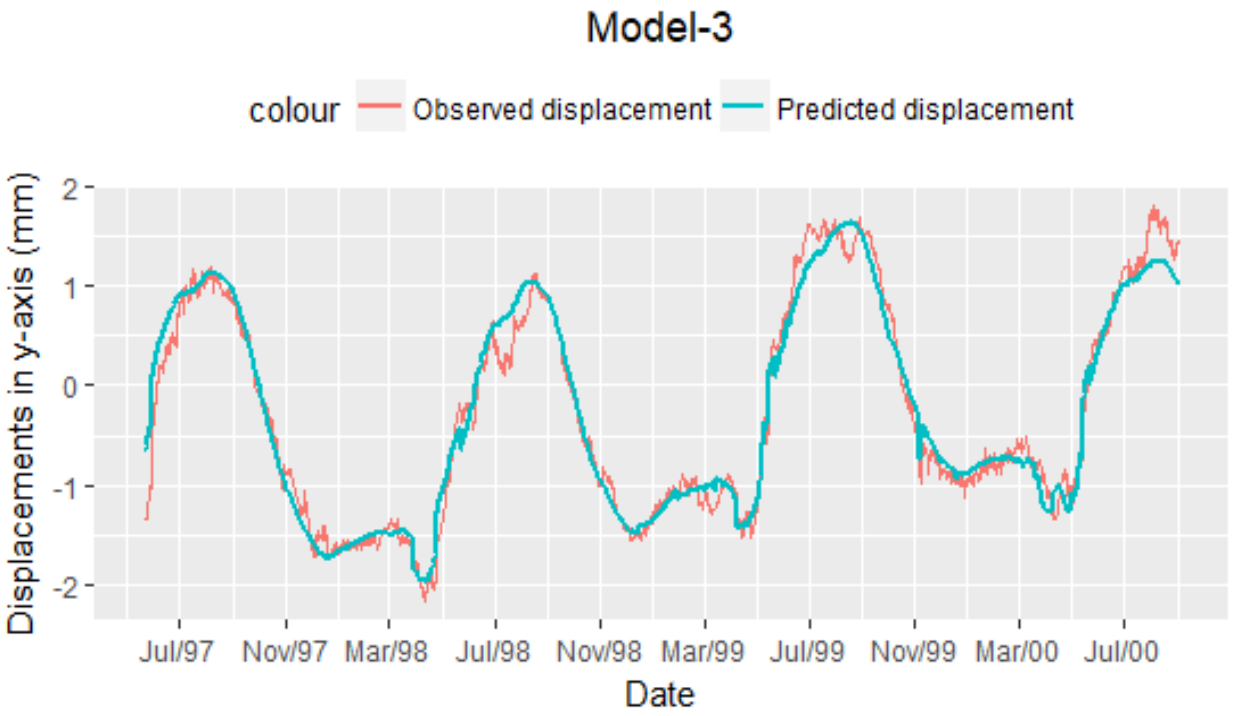

Figure 4.44 - Comparison between observed and predicted displacements for model-3 in y-axis

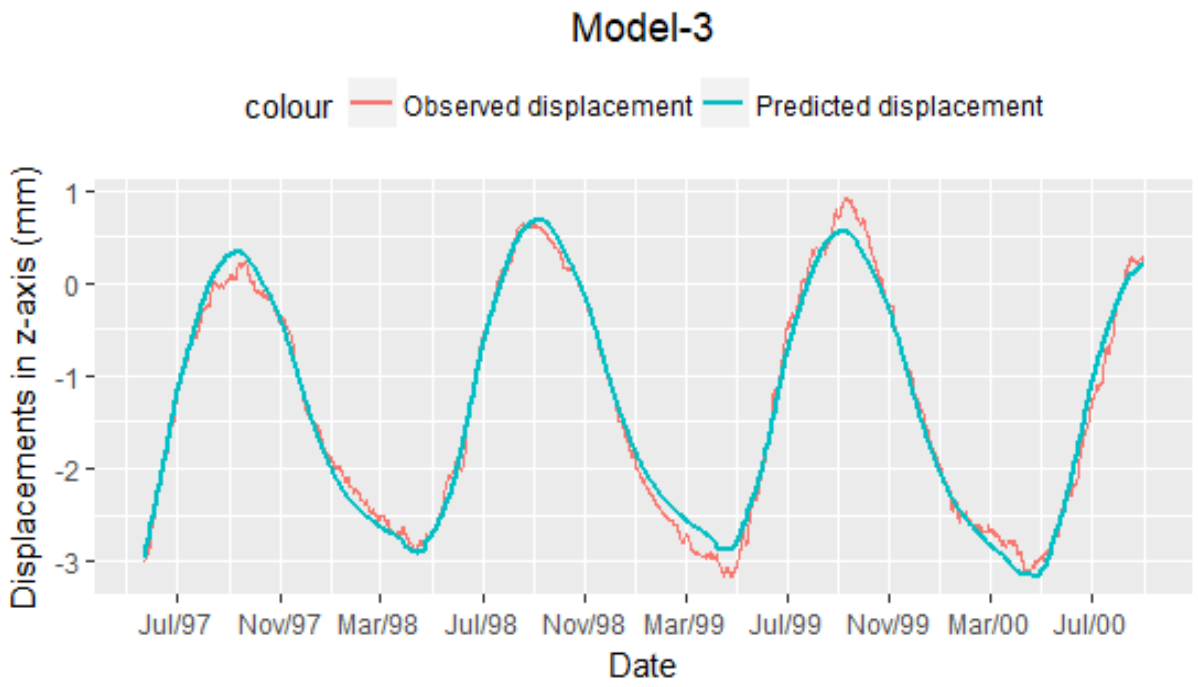

Figure 4.45 - Comparison between observed and predicted displacements for model-3 in z-axis

Observing the above plots, we can say that the non-parametric model is more flexible around the fluctuations and but can capture them marginally better than the two other models. This is because the non-parametric seasonal curve is an estimate of the all the seasons of the time series. It averages out the fluctuations in all the seasons which might not improve the fit for any particular year but will reduce the overall errors in prediction

Residual diagnostics were performed by plotting residual plots as a function of the fitted displacements to check whether the model residuals were white noise or not. Figure 4.46, 4.47, and 4.48 depicts the residuals of model-3 in x, y, and z-axis. Comparing figure 4.46 with figure 4.21 & 4.34 shows that the residual curve for model-3 in x-axis has a lesser trend than the residual curve for model-1 and model-2 in the x-axis. This is because the non-parametric model captures the daily air temperature fluctuations better than model-1 and model-2. The residual plot is still not entirely white noise, but it is because the seasonal curve obtained from the Loess fit was a

yearly seasonal curve which was used for all other years. There can be inter-annual seasonal changes which cannot be captured by this model.

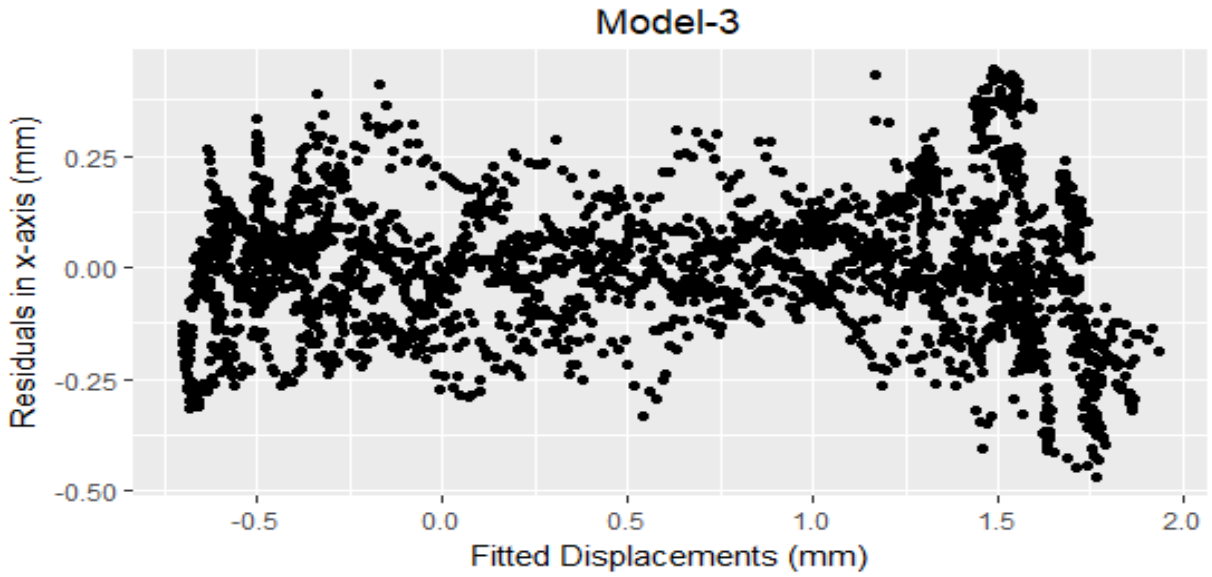

Figure 4.46 - Residual plot for model-3 as a function of the fitted displacements in x-axis

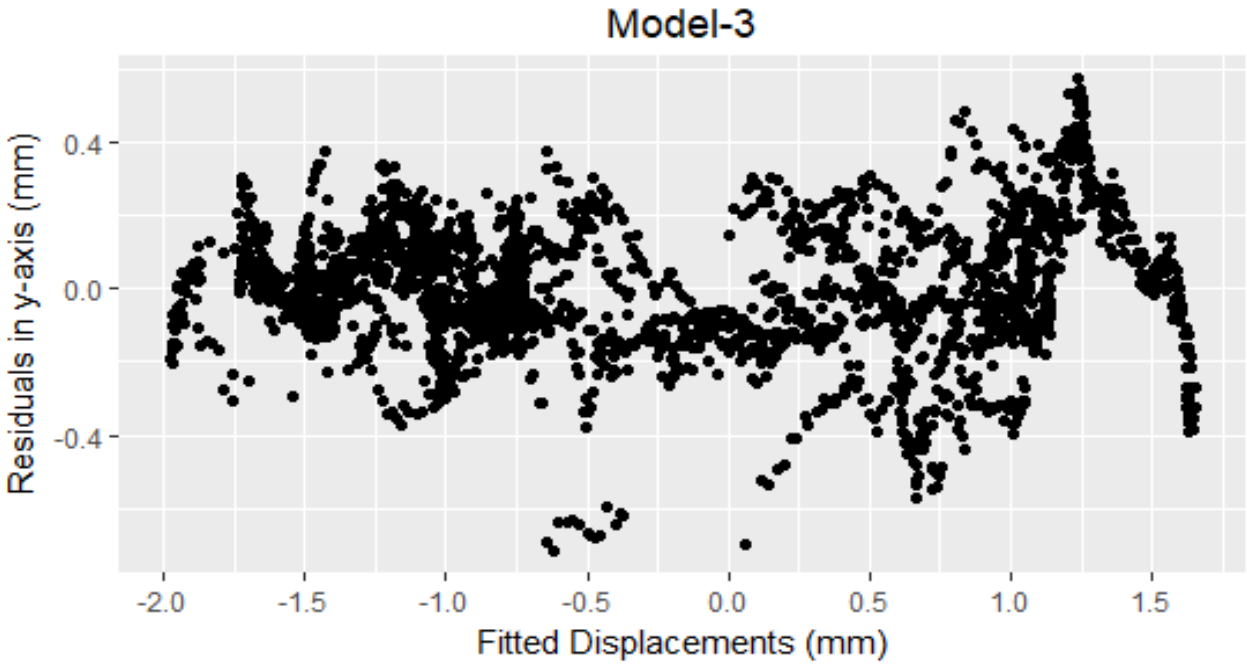

Figure 4.47 - Residual plot for model-3 as a function of the fitted displacements in y-axis

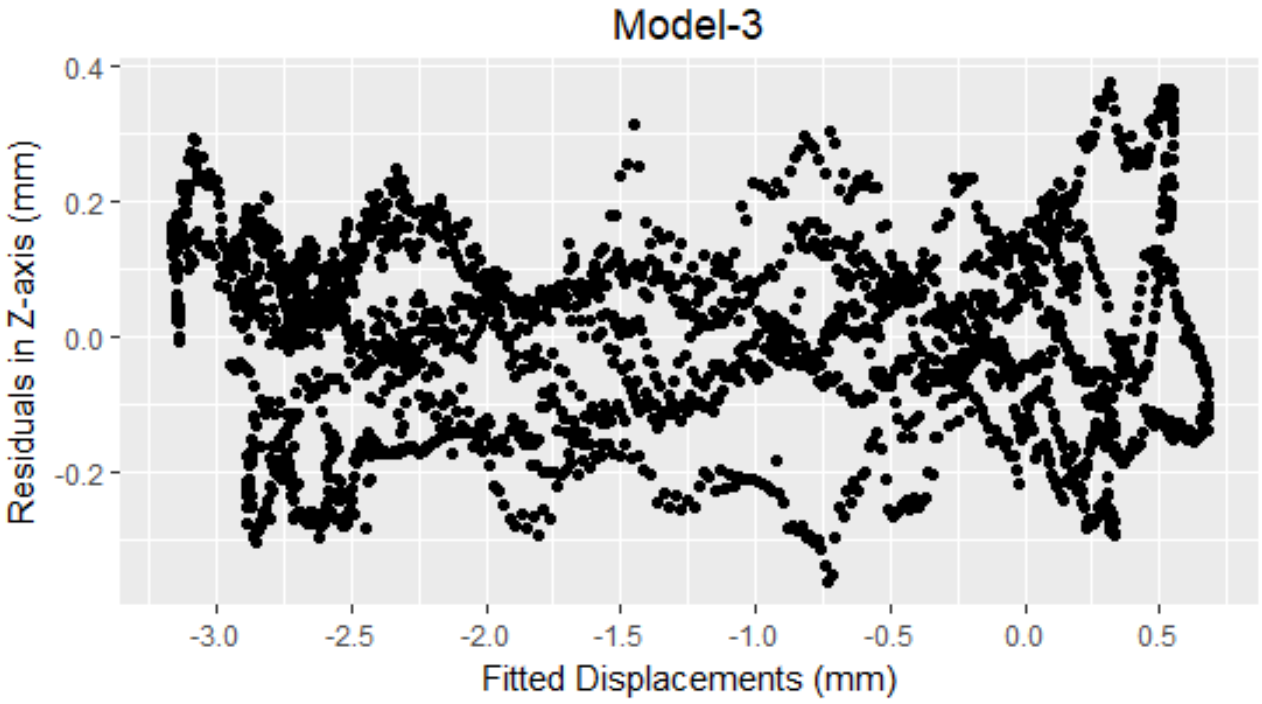

Figure 4.48 - Residual plot for model-3 as a function of the fitted displacements in z-axis

The residual plot for model-3 in the y-axis is very similar to the two other models, but the residual plot for model-3 in z-axis shows less autocorrelation than the ones in model-1 and model-2. It can be concluded that model-3 reduces the residuals, the trend in the residual plots and fits the fluctuations caused due to daily temperature marginally better than the other two models.

4.3.4 Model – 4 (Hydrostatic Seasonal Lagged Temperatures Time model)

Air temperature is an external load which causes delayed displacements in dams, therefore, using lagged air temperature data as independent variables should improve the model and reduce the autocorrelation in the residual analysis. In this model both trigonometric functions and lagged air temperature variables were used to capture the fluctuations more accurately. 18 different lagged variables of the air temperature (named Tempfit) were selected, each with a lag of 10 observations (5days). Therefore three months of lagged air temperature was selected for the model. The seasonal term was multiplied by a linear term of time similar to the seasonal term in model-2.

Before preparing the lagged air temperature variables, it is essential to filter the noise from the temperature signal. To forecast future displacements, we need a smooth curve of the air temperature so that it can itself be forecasted with lower uncertainties. Therefore, a locally weighted linear fit was prepared for the temperature data. The span selected was 0.15 for a total of 2213 observations. Figure 4.48 depicts both the observed air temperature data as well as the filtered signal. A higher span increases the smoothness of the filtered curve and vice versa.

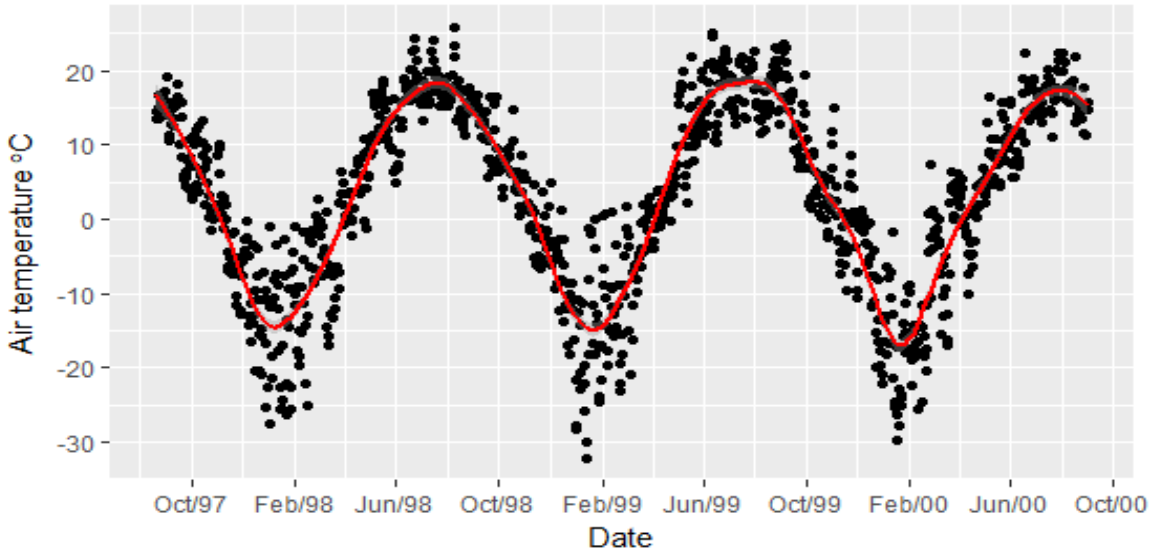

Figure 4.49 - Filtered fit of the air temperature data

The fitted temperature curve is then used to prepare lagged variables of temperature and can be used for future predictions as well.

Tables 4.7, 4.8, & 4.9 depict the regression summary of model-4.

Variables	Coefficients	Standard Error	t-value	P-value (>|t|)
Intercept	-3.213e+00	4.875e-02	-65.902	< 2e-16
Z	8.024e-02	2.076e-02	3.864	0.000115
t	1.347e-03	1.301e-04	10.355	< 2e-16
t2	-3.245e-06	2.304e-07	-14.085	< 2e-16
t3	1.975e-09	1.264e-10	15.627	< 2e-16
Tempfit	4.090e-01	3.488e-02	11.723	< 2e-16
Tempfit10	-5.450e-01	7.153e-02	-7.619	3.79e-14
Tempfit20	3.809e-01	5.745e-02	6.631	4.19e-11
Tempfit40	-6.085e-01	7.928e-02	-7.676	2.46e-14
Tempfit50	9.292e-01	1.350e-01	6.881	7.76e-12
Tempfit60	-7.254e-01	1.215e-01	-5.969	2.78e-09
Tempfit70	2.317e-01	9.274e-02	2.499	0.012530
Tempfit80	4.902e-01	1.284e-01	3.818	0.000138
Tempfit90	-7.417e-01	1.528e-01	-4.853	1.30e-06
Tempfit100	6.276e-01	1.019e-01	6.159	8.69e-10
Tempfit120	-5.233e-01	9.450e-02	-5.538	3.43e-08
Tempfit130	7.337e-01	1.268e-01	5.786	8.23e-09
Tempfit140	-5.773e-01	7.727e-02	-7.472	1.14e-13
Tempfit160	3.904e-01	5.705e-02	6.844	9.99e-12
Tempfit170	-4.665e-01	7.106e-02	-6.565	6.47e-11
Tempfit180	3.575e-01	3.445e-02	10.375	< 2e-16
sin (T)	-2.896e+00	1.093e-01	-26.504	< 2e-16
cos (T)	-7.909e-02	5.201e-02	-1.521	0.128476
sin (T).cos(T)	7.098e-02	2.050e-02	3.462	0.000547
$\sin^2$ (T)	9.340e-01	2.195e-02	42.555	< 2e-16
t*sin (T)	9.367e-05	1.372e-05	6.830	1.10e-11
t*cos (T)	5.669e-04	2.323e-05	24.407	< 2e-16
t*sin (T)*cos (T)	-4.485e-04	2.150e-05	-20.859	< 2e-16
t*$\sin^2$ (T)	-3.835e-05	2.100e-05	-1.826	0.067939
Residual standard error			0.1012 on 2184 degrees of freedom	
Multiple R-squared			0.9936	
Adjusted R-squared			0.9935	
Predicted R-squared			0.9934	

Table 4.7 - Regression summary of model-4 in x-axis

Variables	Coefficients	Standard Error	t-value	p-value (|t|>0)
Intercept	-1.373e+00	6.624e-02	-20.727	< 2e-16
Z	1.461e+00	2.887e-02	50.596	< 2e-16
t	-5.525e-03	1.810e-04	-30.532	< 2e-16
t2	1.234e-05	3.238e-07	38.104	< 2e-16
t3	-6.854e-09	1.773e-10	-38.662	< 2e-16
Tempfit10	-4.176e-02	1.930e-02	-2.164	0.030597
Tempfit30	1.115e-01	6.283e-02	1.775	0.076022
Tempfit40	-1.586e-01	1.045e-01	-1.518	0.129144
Tempfit50	2.376e-01	1.177e-01	2.019	0.043636
Tempfit60	-3.358e-01	1.086e-01	-3.092	0.002014
Tempfit70	1.978e-01	6.261e-02	3.160	0.001600
Tempfit100	2.389e-01	6.491e-02	3.680	0.000239
Tempfit110	-3.126e-01	9.845e-02	-3.175	0.001518
Tempfit120	1.885e-01	8.704e-02	2.166	0.030402
Tempfit130	-2.727e-01	6.004e-02	-4.541	5.91e-06
Tempfit150	1.122e-01	3.307e-02	3.392	0.000707
Tempfit170	1.553e-01	4.432e-02	3.503	0.000469
Tempfit180	-1.550e-01	3.093e-02	-5.010	5.87e-07
sin (T)	1.115e+00	1.442e-01	7.734	1.58e-14
sin (T).cos (T)	4.930e-02	1.684e-02	2.928	0.003448
$\sin^2$ (T)	5.837e-01	3.072e-02	18.996	< 2e-16
t*sin (T)	1.698e-04	1.651e-05	10.288	< 2e-16
t*cos (T)	2.078e-04	3.125e-05	6.650	3.68e-11
t*$\sin^2$ (T)	4.947e-04	2.942e-05	16.818	< 2e-16
Residual standard error			0.1451 on 2189 degrees of freedom	
Multiple R-squared			0.9813	
Adjusted R-squared			0.9811	
Predicted R-squared			0.9808	

Table 4.8 - Regression summary of model-4 in y-axis

Variables	Coefficients	Standard Error	t-value	p-value (|t|>0)
Intercept	1.991e+00	3.458e-02	57.566	< 2e-16
Z	1.225e-01	2.175e-02	5.632	2.01e-08
t	9.271e-04	3.521e-05	26.332	< 2e-16
t3	-8.231e-10	2.635e-11	-31.234	< 2e-16
Tempfit	-4.058e-01	3.661e-02	-11.085	< 2e-16
Tempfit10	5.981e-01	7.543e-02	7.928	3.51e-15
Tempfit20	-4.530e-01	6.062e-02	-7.473	1.13e-13
Tempfit40	5.637e-01	8.216e-02	6.861	8.90e-12
Tempfit50	-7.546e-01	1.353e-01	-5.579	2.72e-08
Tempfit60	5.202e-01	1.016e-01	5.120	3.33e-07
Tempfit80	-6.239e-01	1.134e-01	-5.503	4.16e-08
Tempfit90	6.719e-01	1.783e-01	3.769	0.000168
Tempfit100	-4.650e-01	1.698e-01	-2.738	0.006239
Tempfit110	-2.585e-01	1.319e-01	-1.960	0.050130
Tempfit120	7.381e-01	1.328e-01	5.560	3.02e-08
Tempfit130	-8.086e-01	1.502e-01	-5.384	8.05e-08
Tempfit140	5.085e-01	1.397e-01	3.640	0.000278
Tempfit150	2.082e-01	1.312e-01	1.587	0.112665
Tempfit160	-5.771e-01	1.187e-01	-4.860	1.26e-06
Tempfit170	5.889e-01	9.317e-02	6.321	3.15e-10
Tempfit180	-4.256e-01	3.959e-02	-10.750	< 2e-16
sin (T)	3.467e+00	1.076e-01	32.229	< 2e-16
cos (T)	-6.124e-01	5.475e-02	-11.185	< 2e-16
sin (T).cos(T)	6.064e-02	2.152e-02	2.817	0.004884
$\sin^2$ (T)	-9.103e-01	2.109e-02	-43.172	< 2e-16
t*sin (T)	-1.466e-04	1.280e-05	-11.457	< 2e-16
t*cos (T)	-4.409e-04	2.250e-05	-19.592	< 2e-16
t*sin (T)*cos (T)	6.414e-05	2.281e-05	2.811	0.004979
t*$\sin^2$ (T)	2.524e-04	2.155e-05	11.713	< 2e-16
Residual standard error			0.1075 on 2184 degrees of freedom	
Multiple R-squared			0.9844	
Adjusted R-squared			0.9842	
Predicted R-squared			0.9840	

Table 4.9 - Regression summary of model-4 in z-axis

The above regression summaries suggest that there is a relation between the lagged air temperature variables and the displacements since they have a low p-value and significant t-values. While comparing the results of the model-4 to model-1, it can be seen that a significant improvement in

the fit was registered for the x-axis where the Adjusted-R-squared increased from 0.9619 to 0.9935, and an almost 40% decrease in the residual standard error was registered. The residual standard error for the y-axis and x-axis also reduce by 20% and 32% respectively, supplemented with an increase in the adjusted R-squared Therefore it can be said that adding the lagged temperature variables improves the fit of the HST model. Figure 4.49, 4.50, and 4.51 depict the plots between the observed displacements and the predicted displacements with a decomposition of the predicted displacements into four components, hydrostatic, seasonal, temperature, and time. This decomposition is done to check the contribution of each component to the overall displacement.

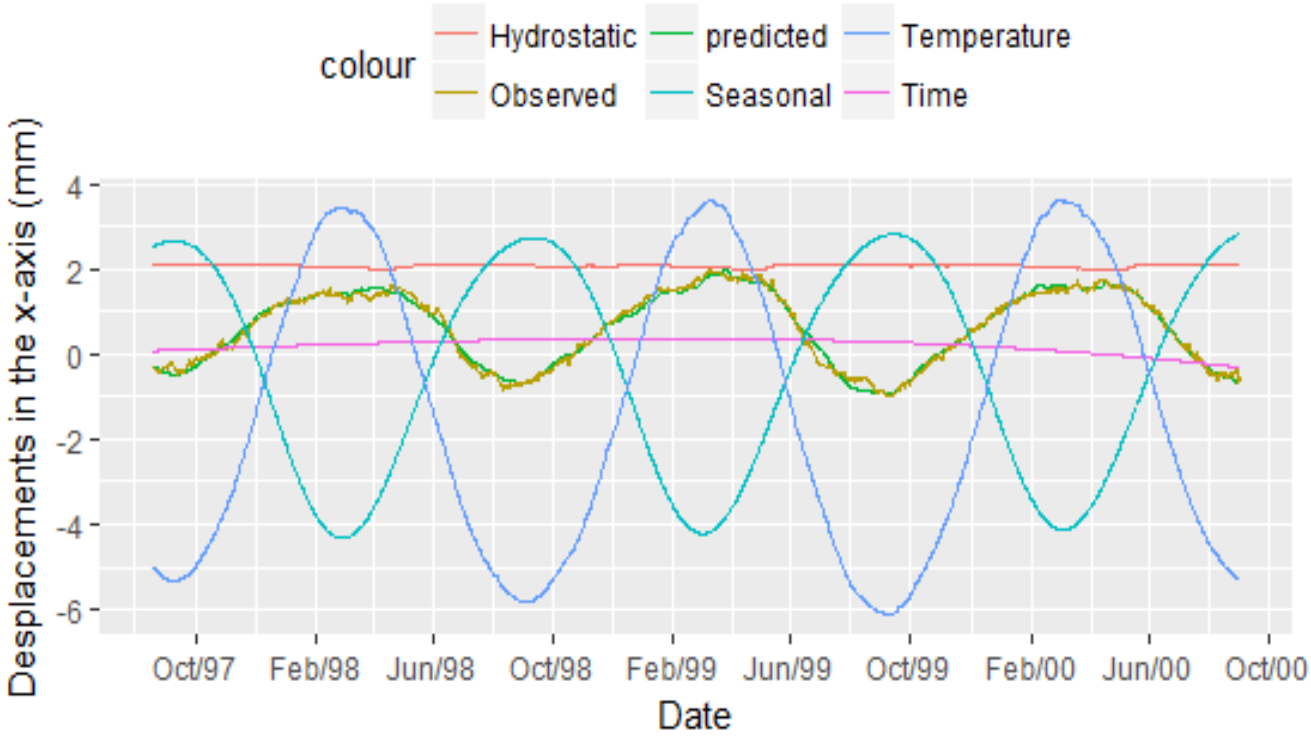

Figure 4.50 - Comparison between the observed and predicted displacements in x-axis

It can be seen that the observed and the predicted curves fit each other very nicely with smaller residuals than the previous three models. The temperature curve lags almost five months behind

the seasonal curve, and both have almost 180-degree phase difference. Model – 4 accounted for the fluctuations more accurately than any other model in the x-axis.

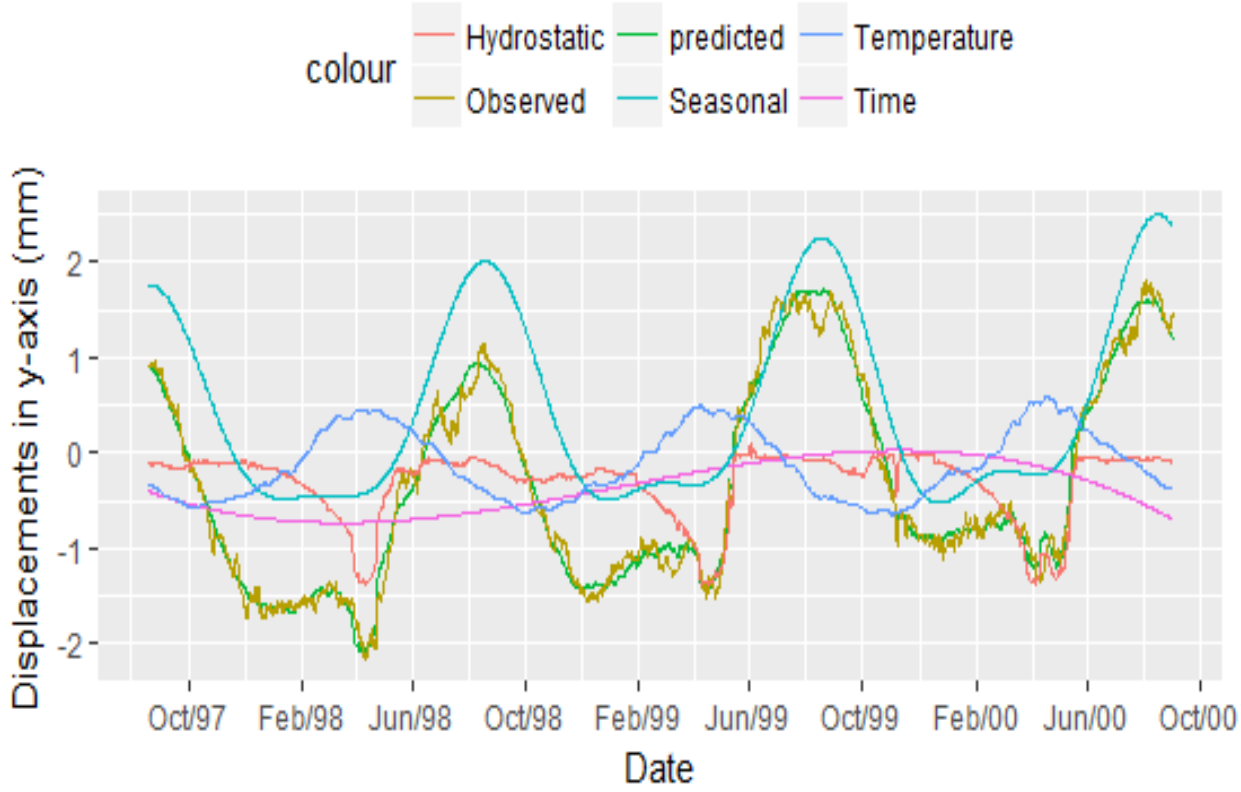

Figure 4.51 - Comparison between the observed and predicted displacements in y-axis

The figure above shows that the fit between the observed and the predicted displacements for the y-axis is good, with the temperature curve lagging behind the seasonal curve by almost five months, similar to the x-axis. Model-4 showed a better fit for the y-axis when compared to any other model.

Figure 4.52 shows that the fit for z-axis is also very accurate and that the lag between the temperature and seasonal model is about six months. It can be seen that the hydrostatic and the time component are more influential in the y-axis and therefore, more deterioration should be observed in the y-axis than other two.

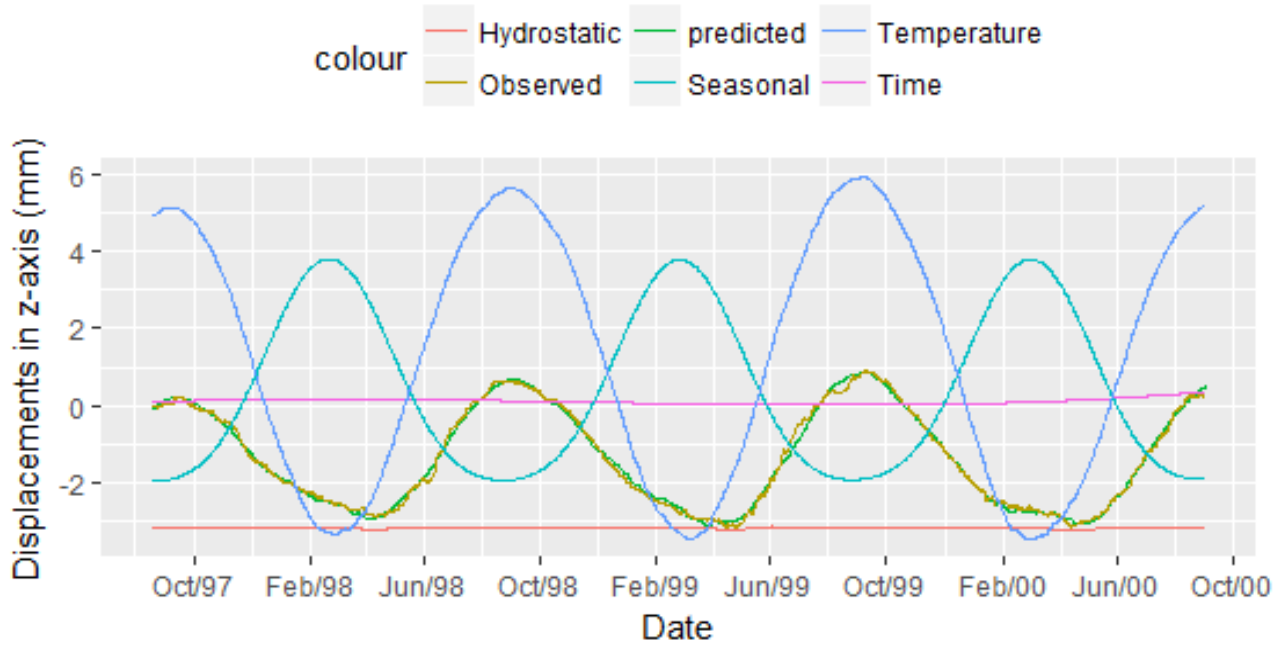

Figure 4.52 - Comparison between the observed and predicted displacements in z-axis

For residual diagnostics: a) The q-q plots and the histograms of the residuals of the displacements in three axes are plotted in the figures 4.53, 4.54, and 4.55.

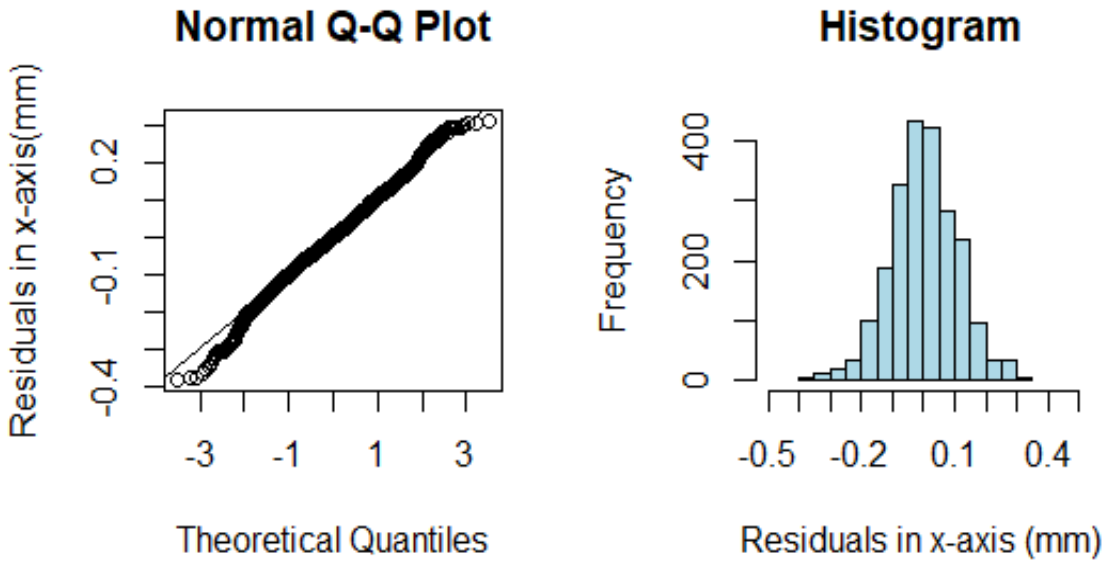

Figure 4.53 - Check for normality of the residuals in x-axis

The residuals for the x-axis seem to be normally distributed since most of the data points lie on the normal distribution line, except some endpoints. Also, the histogram looks like a normal distribution case. Therefore, it can be said that the assumptions of normal residuals holds true in the model for the x-axis.

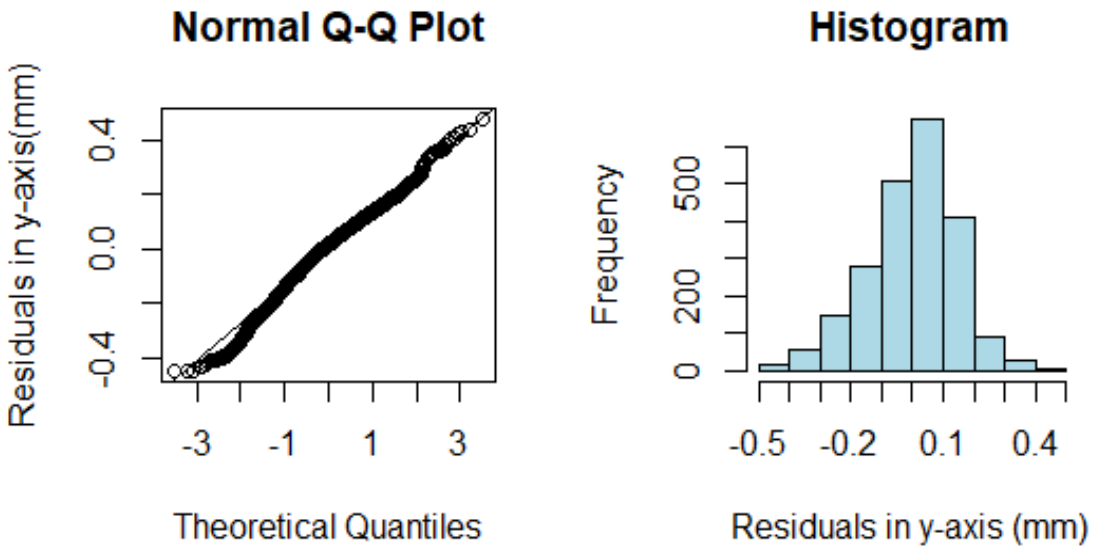

Figure 4.54 - Check for normality of the residuals in y-axis

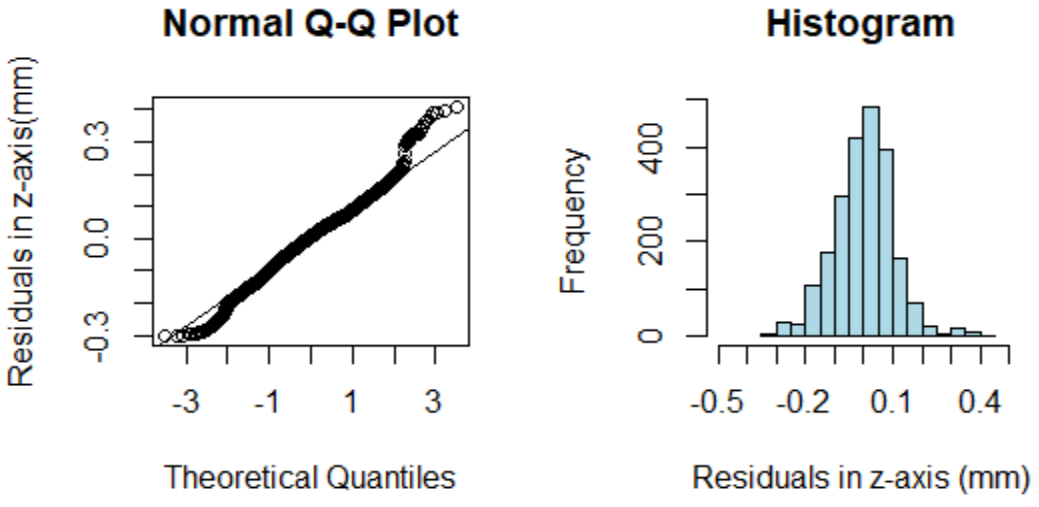

Figure 4.55 - Check for normality of the residuals in z-axis

From the above q-q plots and histograms it can be seen that the residuals in the y-axis are normal whereas the residuals in the z-axis do not follow normal distribution at the ends. However, it is approximately normal.

b) The residual plots as a function of the fitted displacement are plotted for all three axes to check for white noise.

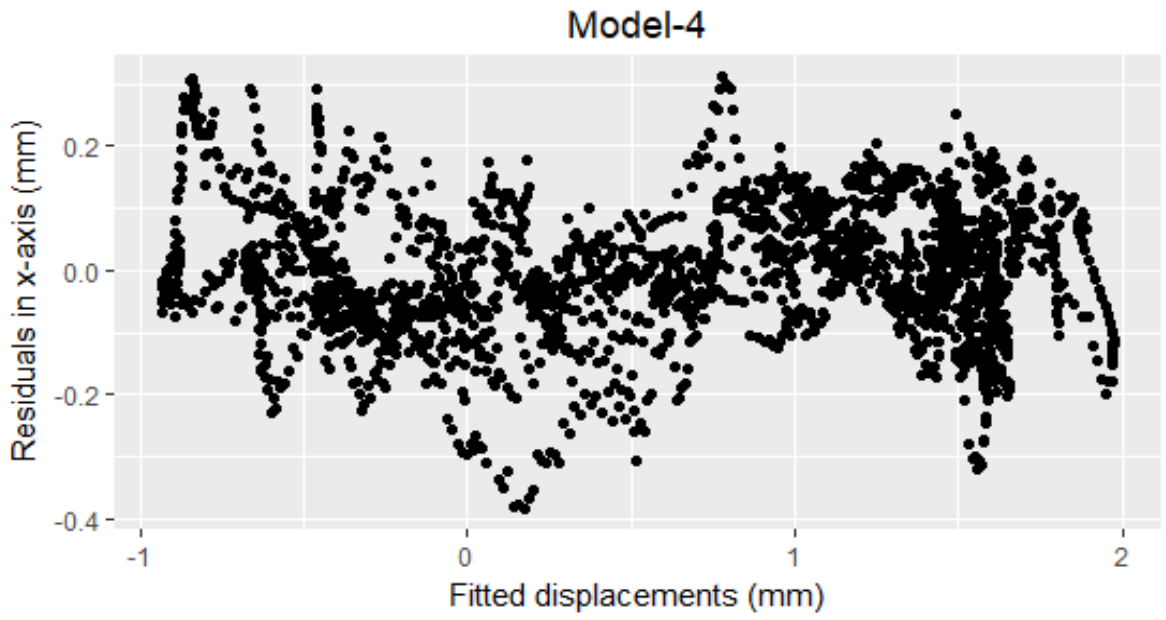

Figure 4.56 - Residual plot as a function of fitted displacement in x-axis

The residuals for model-4 in the x-axis are smaller in magnitude than the other models, and the autocorrelation of the residuals has also decreased when compared to model-1 and model-2.

The residual plot for model-4 in the y-axis also shows improvements over other models. The shape of the residual remains almost the same, but there was a reduction in the autocorrelation of the residuals. Also, the overall magnitude of the residuals decreased.

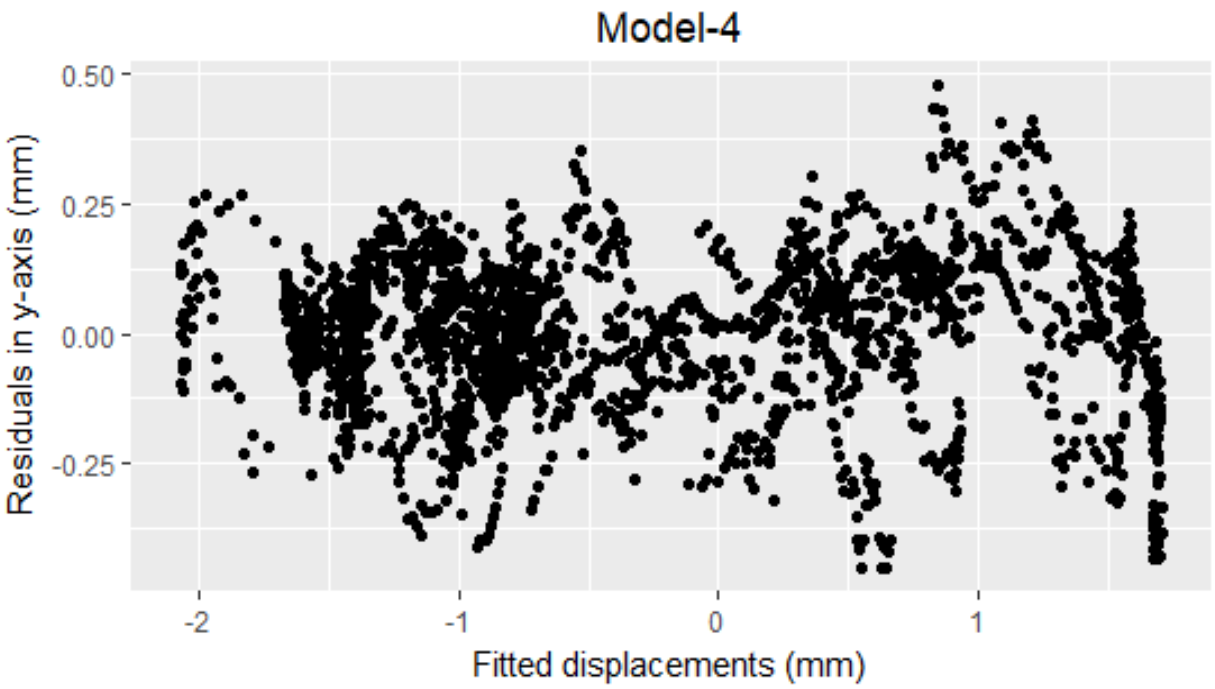

Figure 4.57 - Residual plot as a function of fitted displacement in y-axis

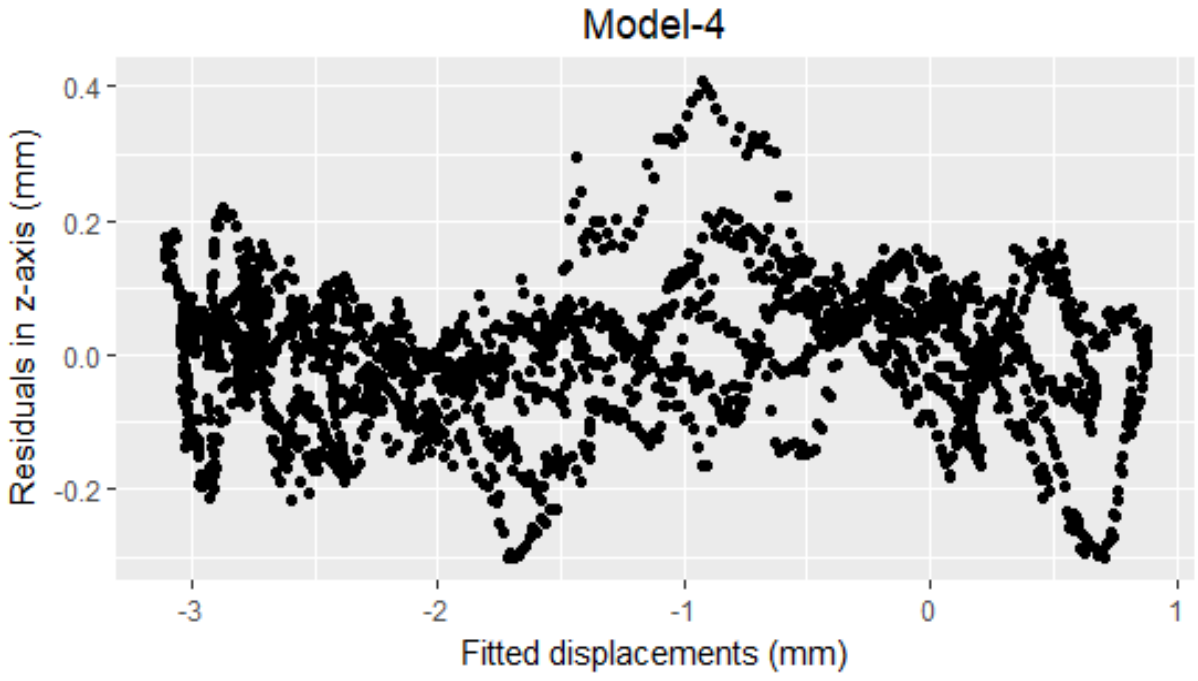

Figure 4.58 - Residual plot as a function of fitted displacement in z-axis

The residual plot for the z-axis shows minor trend and also has a smaller magnitude than the ones from the other models, but still isn't entirely white noise.

It can be concluded that model-4 can account for some of the daily air temperature fluctuations and helps to reduce the scattering in the residual model by as much as 40%. This shows improvement in fit as compared to the other three models. The only disadvantage is that it increases the number of independent variables used in the regression analysis which might overfit the data if the number of observations is low.

5 Statistical analysis of dam piezometer data

5.1 Introduction

As discussed in section 2.2.4 water level variations in the reservoir and rainfall have been discussed as essential predictors to forecast piezometer levels in an arch dam. In this study, displacements of the dam in radial, tangential & vertical axis were used as additional predictors to improve the fit of two piezometers, one on the upstream end (PZP5m) and another on the downstream (PZP5v). The description of the monitoring data is presented in section 3.3.5.

Displacements in a dam cause small cracks on the surface of the dam which deteriorate with time. At high pore pressures, these cracks open and lead to more seepage, which can suddenly increase the displacement of the dam. The stability of the dam can be questioned if there is a sudden increase in the pore pressure at peak displacements. Therefore the model uses displacements of the dam to improve the predictions of the piezometer model. Rainfall effects and the delay of response to the changes in water level were neglected.

5.2 Establishing relationships between the variables

Scatterplots between the response variable (piezometer head) and the independent variables (reservoir level & displacements) were plotted to understand the relationships between them better. Figures 5.1 & 5.2 show the relationship between the reservoir level and the piezometer levels in the upstream and downstream directions.

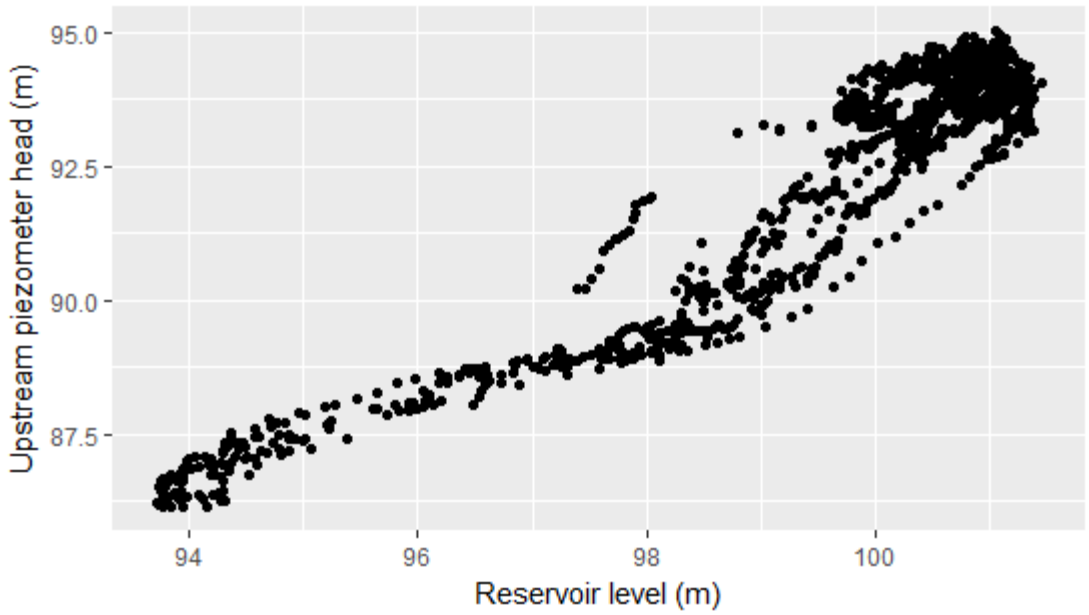

Figure 5.1 - Relationship between the upstream piezometer level and water level in the reservoir

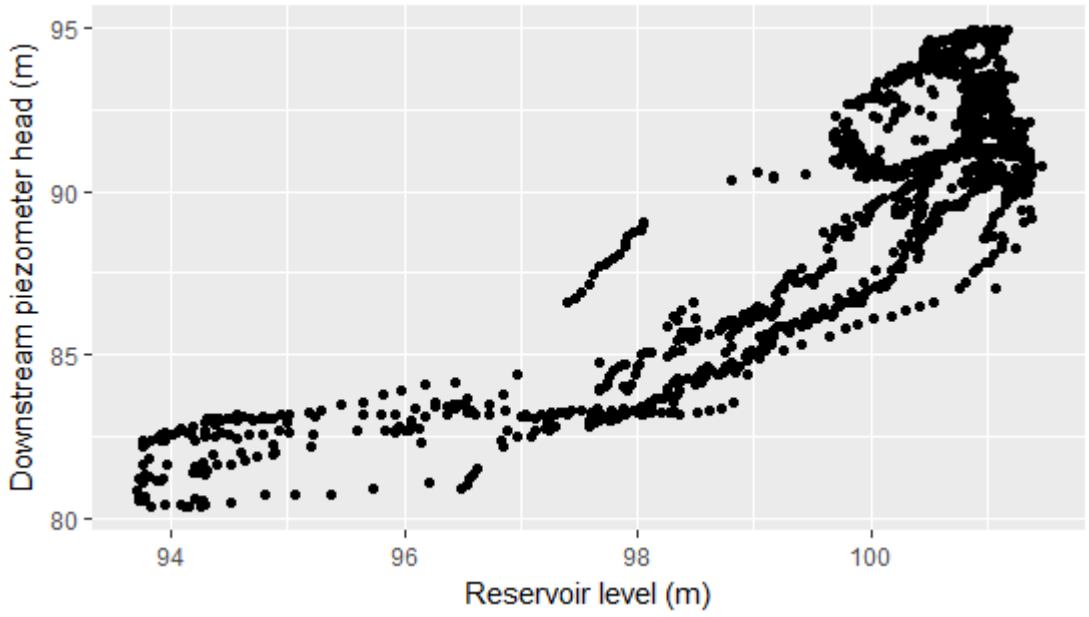

Figure 5.2 - Relationship between the downstream piezometer level and water level in the reservoir

It can be seen that there is some hysteresis in the piezometer head with increasing and decreasing reservoir levels, but it is not very significant and can be accounted for approximately with the help

of a polynomial function of the reservoir level. The upstream piezometer is found to have less hysteresis than the one on downstream.

Figure 5.3, 5.4 and 5.5 show the relationship between the upstream piezometer levels and the displacement of the pendulum in x, y & z-axis.

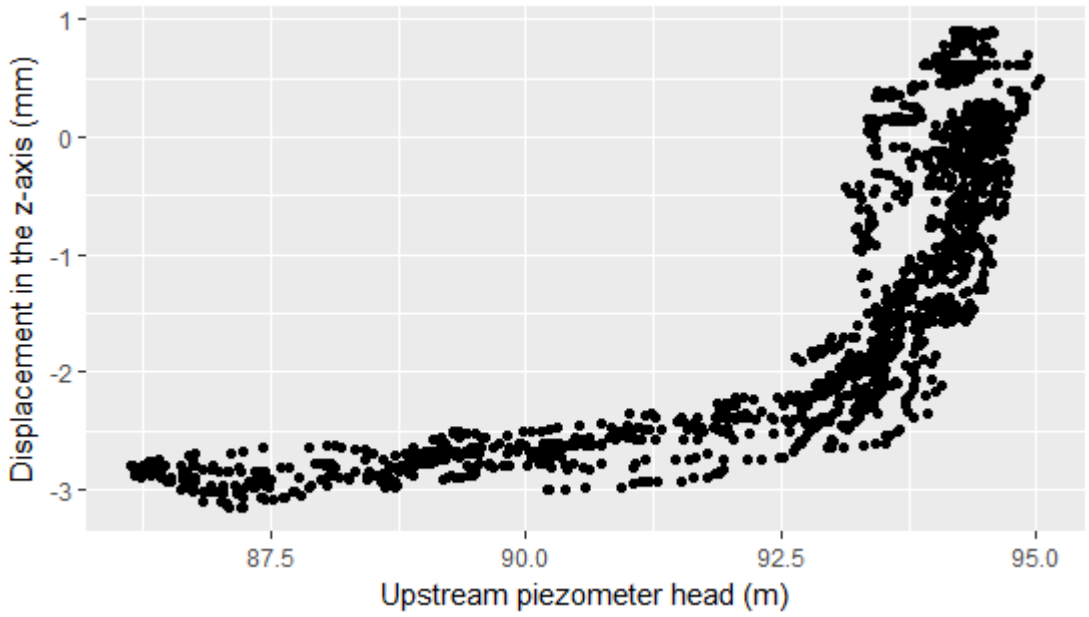

Figure 5.3 - Relationship between the upstream piezometer head and displacement in the z-axis

The relationship between the upstream piezometer head and the displacement in the z-axis is non-linear because after a particular piezometer head, here at 93.75m, the cracks open up and leads to a sudden increase in the displacements. This non-linear relationship is modeled better with the help of segmented regression. The piezometer head was fitted piecewise with an estimated breakpoint at Dz = -2.5mm.

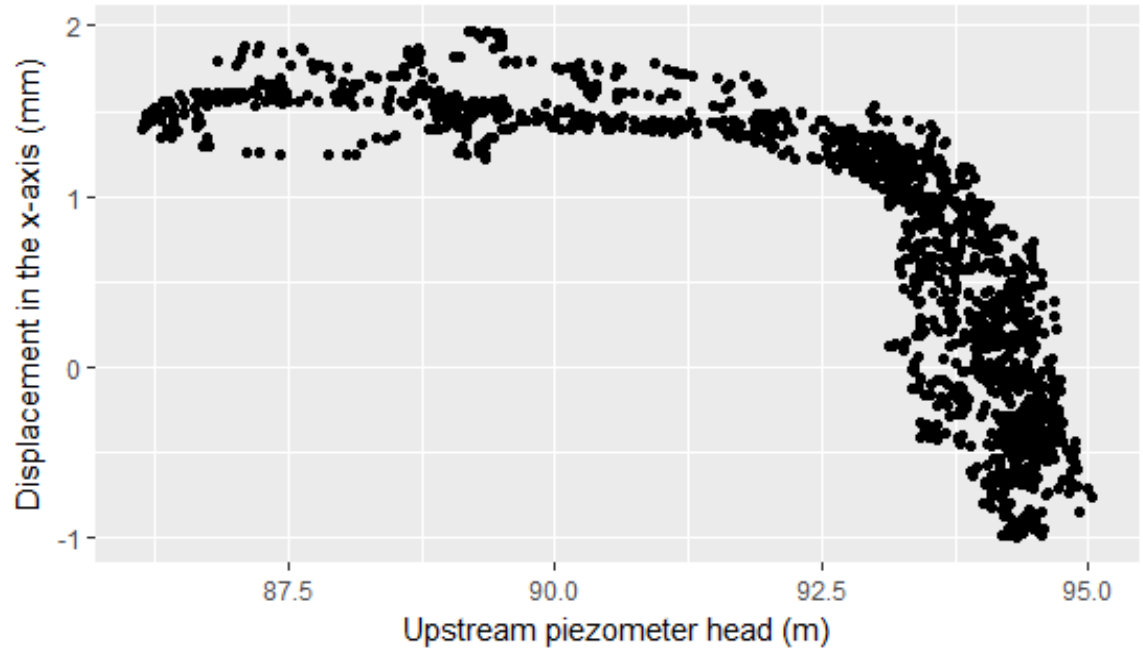

Figure 5.4 - Relationship between the upstream piezometer head and displacement in the x-axis

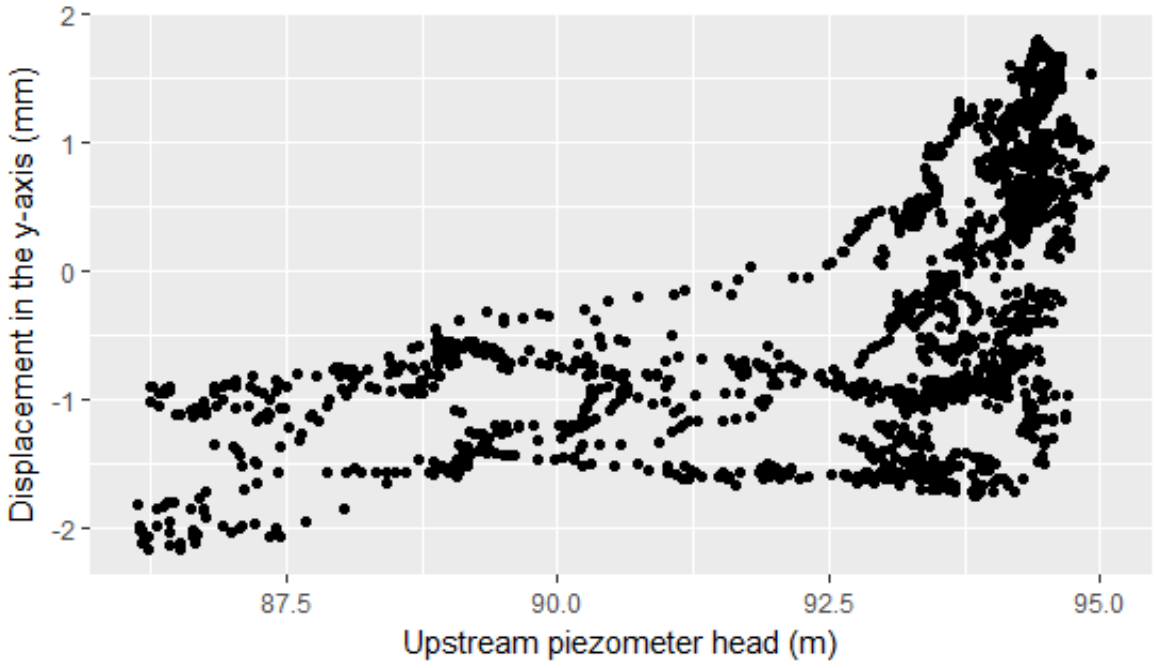

Figure 5.5 - Relationship between the upstream piezometer head and displacement in the y-axis

Similarly, the relation between the upstream piezometer head and the displacements in the x, and the y-axis is also non-linear. However, there is a high degree of hysteresis in the piezometer head due to upstream and downstream displacements in the y-axis. Therefore segmented fit of the piezometer head was prepared for the variables Dx & Dz only, and Dy was not used as apredictor variable in this model. The breakpoint selected for x-axis was at Dx = 1.

Figures 5.6, 5.7 and 5.8 show the relationship between the downstream piezometer levels and the displacement of the pendulum in x, y & z-axis

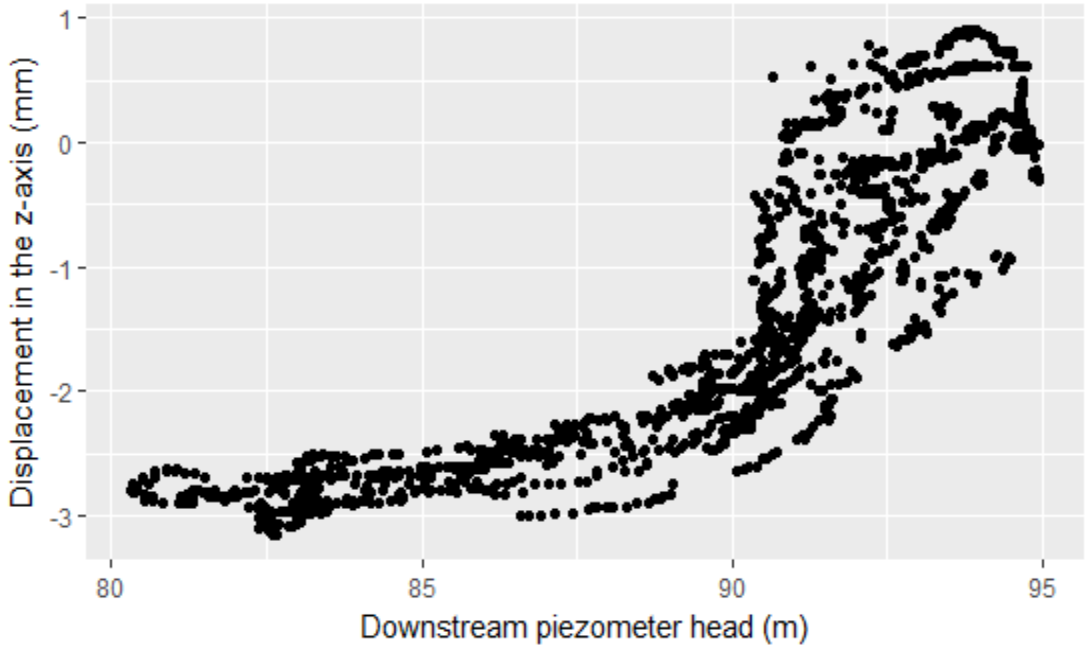

Figure 5.6 - Relationship between the downstream piezometer head and displacement in the z-axis

The relationship between the downstream piezometer level and the displacements in the x, y, & z-axes is non-linear but has more hysteresis than the one upstream. A segmented fit for the variable Dz was fit at Dz = -2 which means that the above figure was fitted using two linear fits, one between Dz = -3 to -2, and the other one between Dz = -2 to 1.

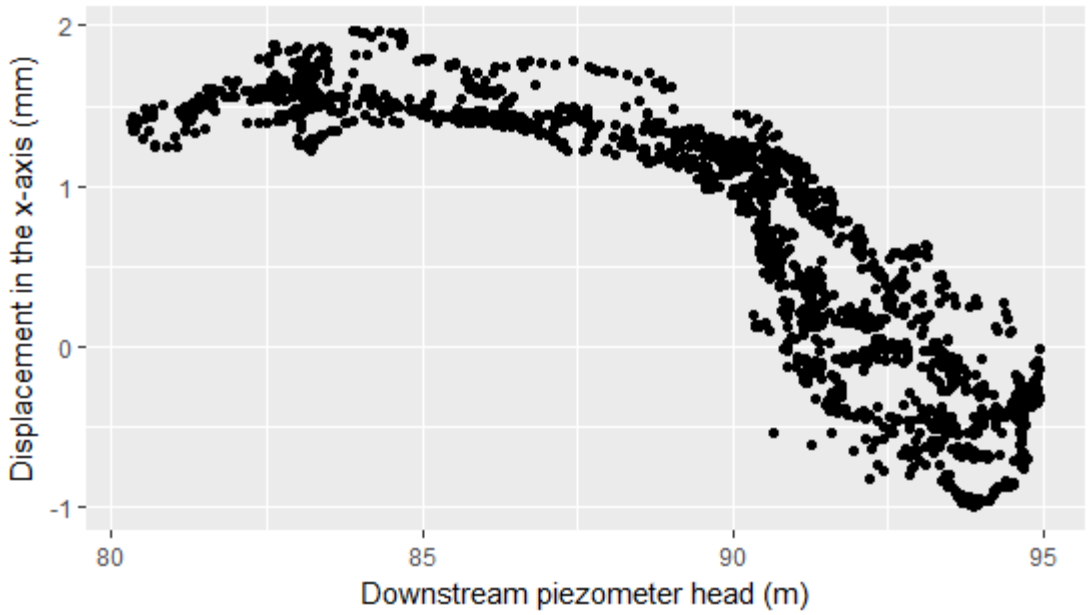

Figure 5.7- Relationship between the downstream piezometer head and displacement in the x-axis

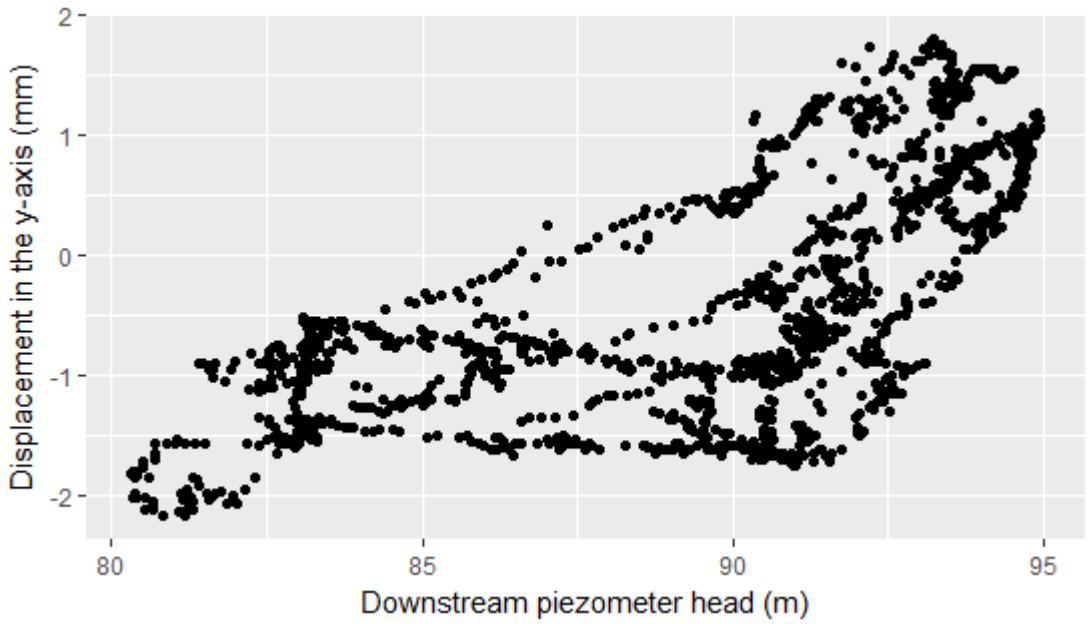

Figure 5.8 - Relationship between the downstream piezometer head and displacement in the y-axis

Similar to the observations made in the upstream piezometer, the relationships between the downstream piezometer head and the displacements in x, y, and z-axis are also non-linear and have

more hysteresis in the y-axis. Therefore, Dy was dropped from the model for the downstream piezometer as well, and Dx & Dz were fitted piecewise.

5.3 Model fitting

Two separate models were made for the two piezometers, one with normal linear fit, and the other with a segmented linear fit for the displacements.

5.3.1 Model-1 - PZP5m represents the normal linear model for the upstream piezometer and PZP5v represents the linear model for the downstream piezometer. The models look like:

PZP5m → $PZm \sim a_1H + a_2H^2 + a_3H^3 + a_4Dx + a_5Dz$ 5.1

PZP5v → $PZv \sim a_1H + a_2H^2 + a_3H^3 + a_4Dx + a_5Dz$ 5.2

Where PZm and PZv are the piezometer levels in the upstream and downstream piezometers, H is the water level in the reservoir and, Dx & Dy are the displacements in x and y-axis respectively. A stepwise regression algorithm was used to compute the coefficients of the predictor variables. Tables 5.1 & 5.2 represent the regression summaries of the piezometers in the upstream and the downstream directions respectively.

Variables	Coefficient estimate	Standard error	t-value	P-value
Intercept	4.340e+02	1.763e+01	24.61	<2e-16
H	-5.713e+00	2.702e-01	-21.14	<2e-16
H^3	2.311e-04	9.394e-06	24.60	<2e-16
Dx	1.112e+00	6.661e-02	16.69	<2e-16
Dz	1.338e+00	4.398e-02	30.42	<2e-16
Residual standard error			0.3866 on 1655 degrees of freedom	
Multiple R-squared			0.9727	
Adjusted R-squared			0.9727	
Predicted R-squared			0.9725	

Table 5.1 - Regression summary of the upstream piezometer

The regression summary shows that all the variables are good predictors of the piezometer levels since the p-values are very low, the predicted R-squared is high, and the residual standard error is low. However, the reservoir level and the displacement in the z-axis (vertical axis) are the dominant predictors.

Independent variables	Coefficient estimate	Standard error	t-value	P-value
Intercept	1.532e+03	6.357e+01	24.106	<2e-16
H	-3.061e+01	1.299e+00	-23.564	<2e-16
H^2	1.621e-01	6.636e-03	24.429	<2e-16
Dx	5.432e-01	1.604e-01	3.386	0.000725
Dz	2.042e+00	1.059e-01	19.280	<2e-16
Residual standard error			0.9325 on 1655 degrees of freedom	
Multiple R-squared			0.9443	
Adjusted R-squared			0.9442	
Predicted R-squared			0.9439	

Table 5.2 - Regression summary of the downstream piezometer

The regression summary for the downstream piezometer showed similar results to the one upstream. Reservoir level and the displacement in the z-axis were the dominant predictors with low p-values. The predicted R-squared was high, and the residual standard error was low. The predicted piezometer levels were compared with the observed piezometer level data to get better idea of the fit of the model, both in upstream and downstream directions. Figures 5.10 & 5.11 depict the comparison of the predicted and the observed piezometer levels for the upstream and the downstream models respectively, by plotting them together as a function of time. Further, regression diagnostics like normality of the residuals, and the residual plots as a function of the fitted piezometer level were checked to validate the model. Normality check is done to satisfy the

assumption of normally distributed residuals with a zero mean and constant variance which is necessary for a valid regression model.

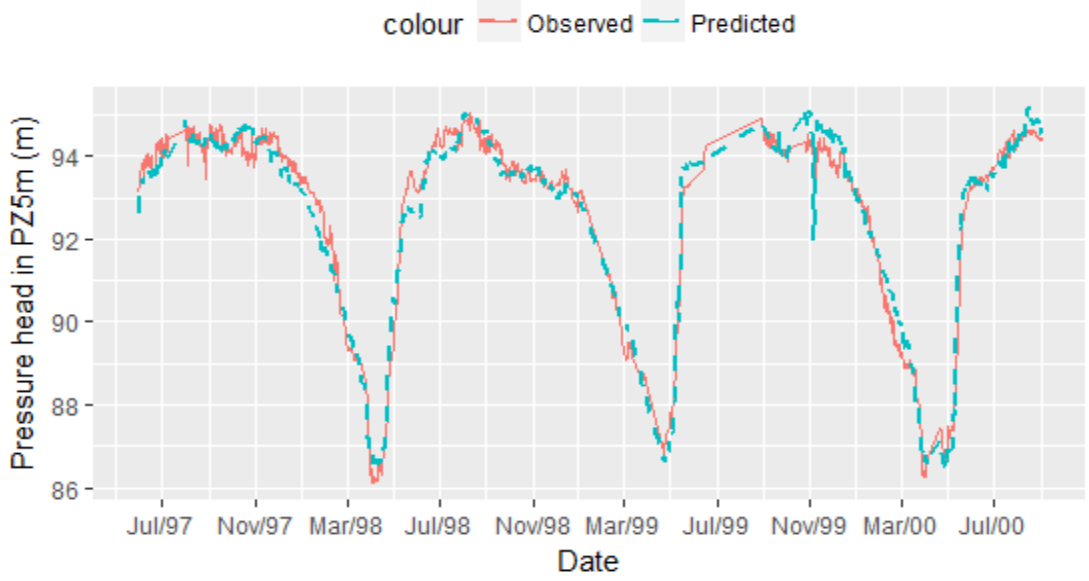

Figure 5.9 - Comparison between the observed and the predicted piezometer levels in the upstream model

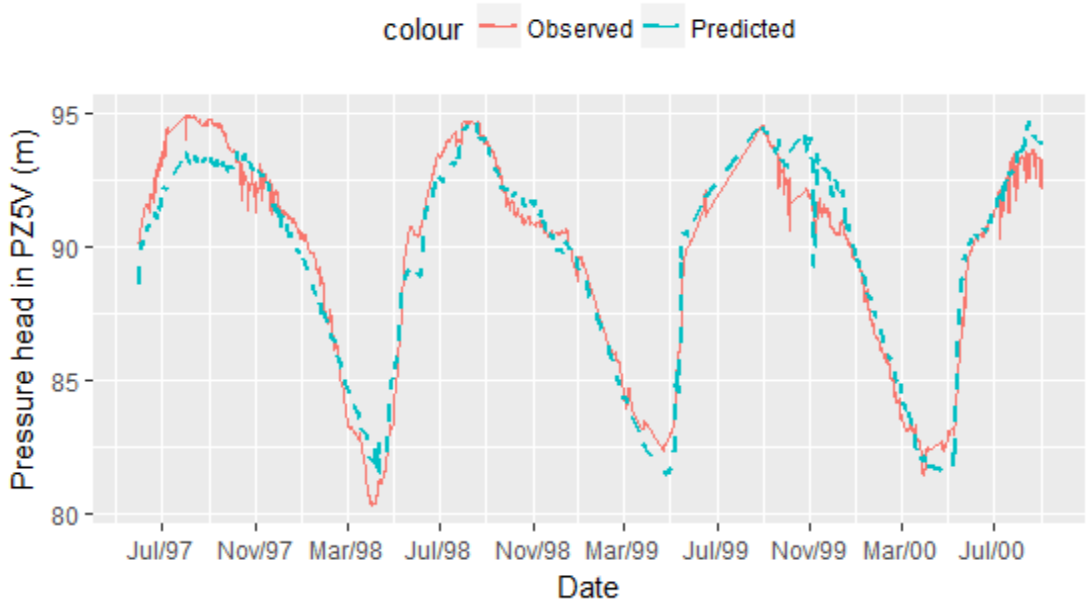

Figure 5.10 - Comparison between the observed and the predicted piezometer levels in the downstream model

The comparison between the observed and the predicted piezometer levels shows that the model performs satisfactorily in the upstream direction but gives larger residuals in the downstream direction. This is due to the presence of higher hysteresis in the data from the downstream piezometer. Figures 5.11 & 5.12 explain the normality of the residuals for the upstream and downstream piezometer models with the help of histograms and q-q plots.

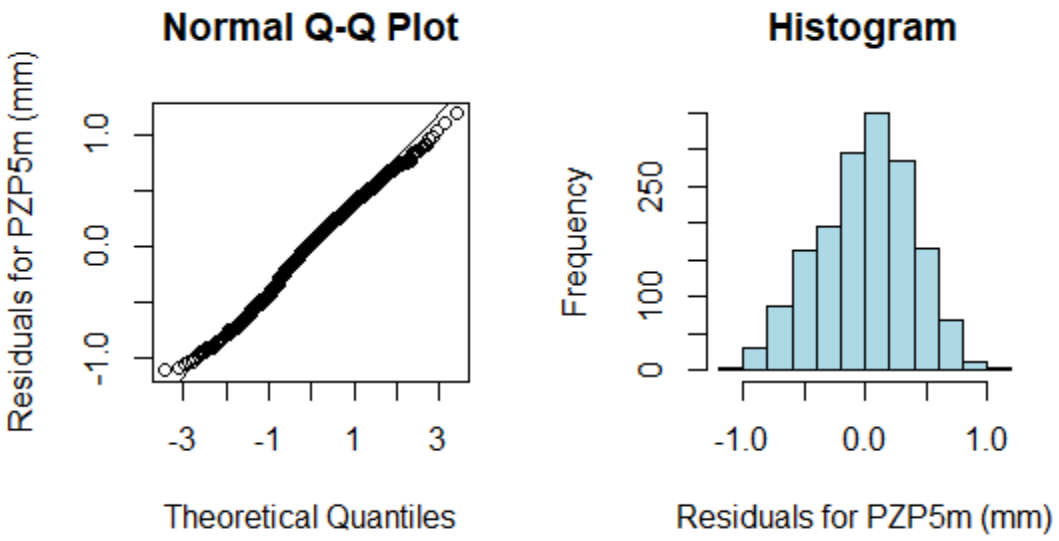

Figure 5.11 - Q-Q plot & histogram of the residuals for the upstream model

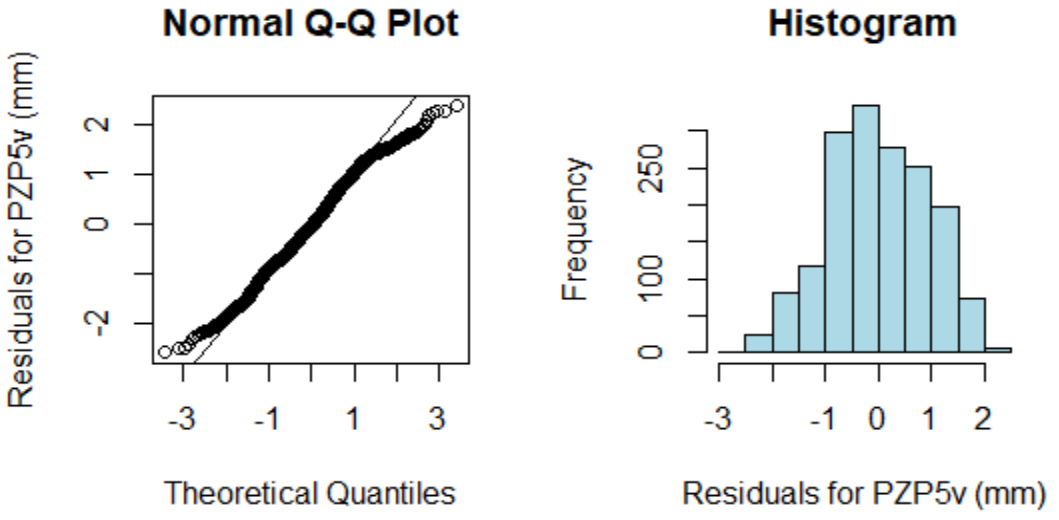

Figure 5.12 - Q-Q plot & histogram of the residuals for the downstream model

The plots for the upstream model suggest that the residuals are normally distributed, but the plots for the downstream model suggest a skewed normal distribution since the points at the ends of the q-q plot do not lie on the theoretical normal distribution line and the histogram also shows the skewed distribution of the residuals. Therefore, it is clear that the downstream gives less accurate results than the upstream model.

5.3.1 Model-2 (segmented model)

Since the relationship between the displacements and the piezometer head is non-linear, a segmented fit of the displacement in the x & z-axis were computed in model-2 while the rest of the model remains the same. Tables 5.3 & 5.4 depict the regression summaries of the segmented upstream and downstream models.

Independent variables	Coefficient estimate	Standard error	t-value	P-value
Intercept	3.310e+02	1.616e+01	20.484	<2e-16
H	-3.951e+00	2.483e-01	-15.907	<2e-16
H^3	1.631e-04	8.692e-06	18.765	<2e-16
Dx	4.287e-01	6.271e-02	6.837	1.14e-11
Dz	3.212e+00	1.143e-01	28.111	<2e-16
U1.Dx	2.221e+00	1.736e-01	12.791	NA
U1.Dz	-2.480e+00	1.203e-01	-20.615	NA
Estimated breakpoint in Dx			1.469	
Estimated breakpoint in Dz			-2.166	
Residual standard error			0.325 on 1651 degrees of freedom	
Multiple R-squared			0.9808	
Adjusted R-squared			0.9807	
Predicted R-squared			0.9805	

Table 5.3 - Regression summary for the segmented upstream model

It can be seen that two new variables named U1.Dx & U1.Dz are introduced in the regression analysis. These variables represent the piecewise fit of the displacement in the x & z-axis, and their

coefficients explain the change in the slope of the linear fit after the estimated breakpoint. The segmented model improved the R-squared values and reduced the overall standard error. The estimated breakpoints in Dx & Dz were 1.469 and -2.166 which could be validated visually by looking at the breakpoints in the scatterplots Figure 5.3 & 5.4.

Independent variables	Coefficient estimate	Standard error	t-value	P-value
Intercept	1.314e+03	6.489e+01	20.254	<2e-16
H	-2.574e+01	1.329e+00	-19.365	<2e-16
H^2	1.360e-01	6.809e-03	19.973	<2e-16
Dx	-9.288e-01	1.624e-01	-5.719	1.27e-08
Dz	5.101e+00	3.038e-01	16.788	<2e-16
U1.Dx	6.466e+00	3.820e-01	16.927	NA
U1.Dz	-4.284e+00	3.226e-01	-13.279	NA
Estimated breakpoint in Dx			1.385	
Estimated breakpoint in Dz			-2.166	
Residual standard error			0.8369 on 1651 degrees of freedom	
Multiple R-squared			0.9553	
Adjusted R-squared			0.9551	
Predicted R-squared			0.9548	

Table 5.4 - Regression summary for the segmented downstream model

Similar to the regression summary of the upstream model, two new variables named U1.Dx & U1.Dz are introduced by the segmented fit, the coefficients of which depict the change in the slope of the fitted line at estimated breakpoints. The segmented downstream model also showed increased R-squared and reduced standard errors (~10%). The reservoir level and the displacement in the z-axis were still the dominant predictors. The estimated breakpoints of the displacements in the x & z-axis were 1.385 and -2.166 which could be validated visually by looking the Figures 5.5 & 5.6.

Further, the piezometer levels predicted by the segmented upstream and downstream models are compared to the observed piezometer levels. Figure 5.15 & 5.16 compare the observed and the

predicted piezometer heads for the upstream and downstream models respectively. It can be seen that the fit for both, the upstream and the downstream models is slightly better than model-1.

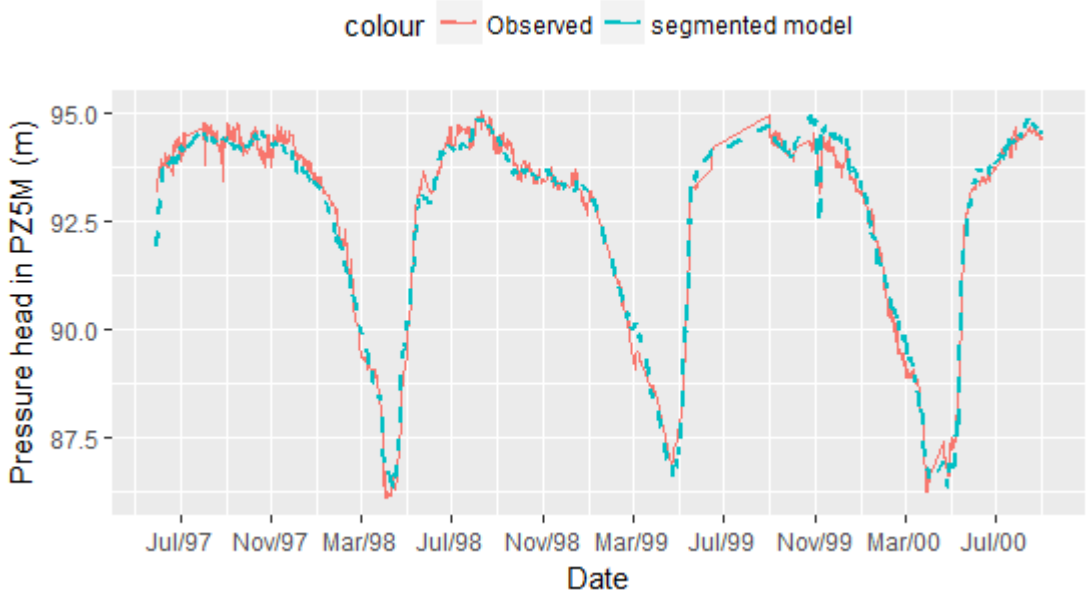

Figure 5.13 - Comparison between the observed and the predicted piezometer levels in the upstream segmented model

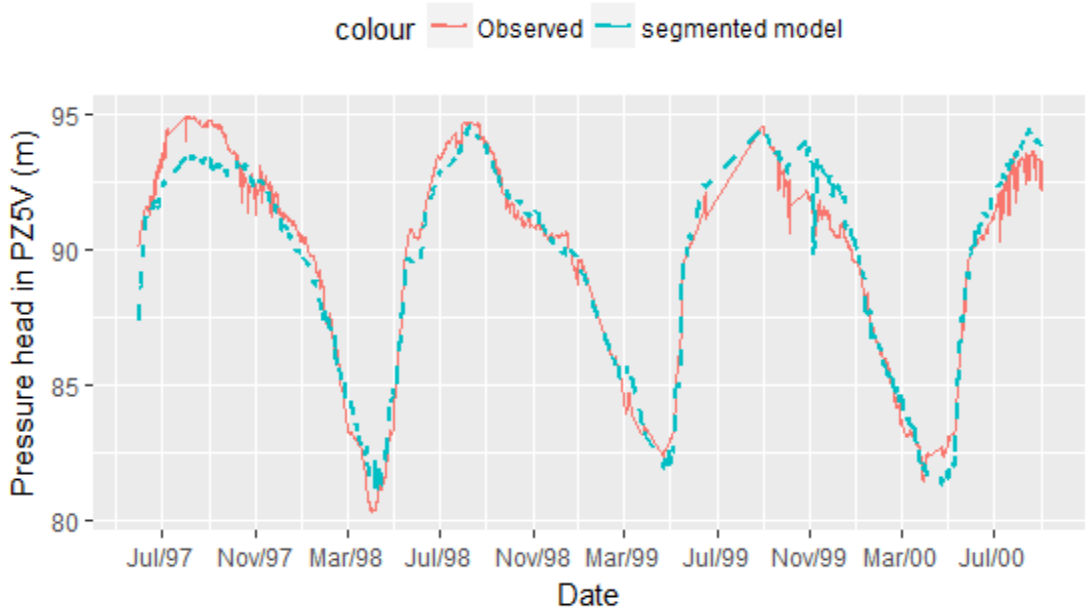

Figure 5.14 - Comparison between the observed and the predicted piezometer levels in the downstream segmented model

The normality of the residuals was checked by plotting histograms and q-q plots. Figures- 5.15 & 5.16 show the histograms and the q-q plots for the upstream and downstream segmented model respectively.

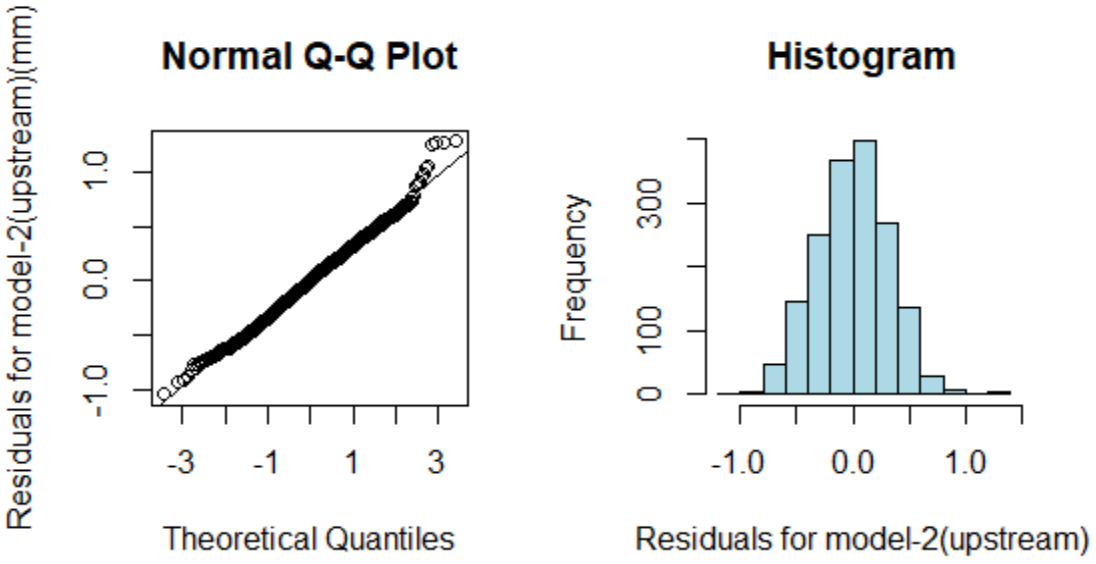

Figure 5.15 - Q-Q plot & histogram of the residuals for the upstream segmented model

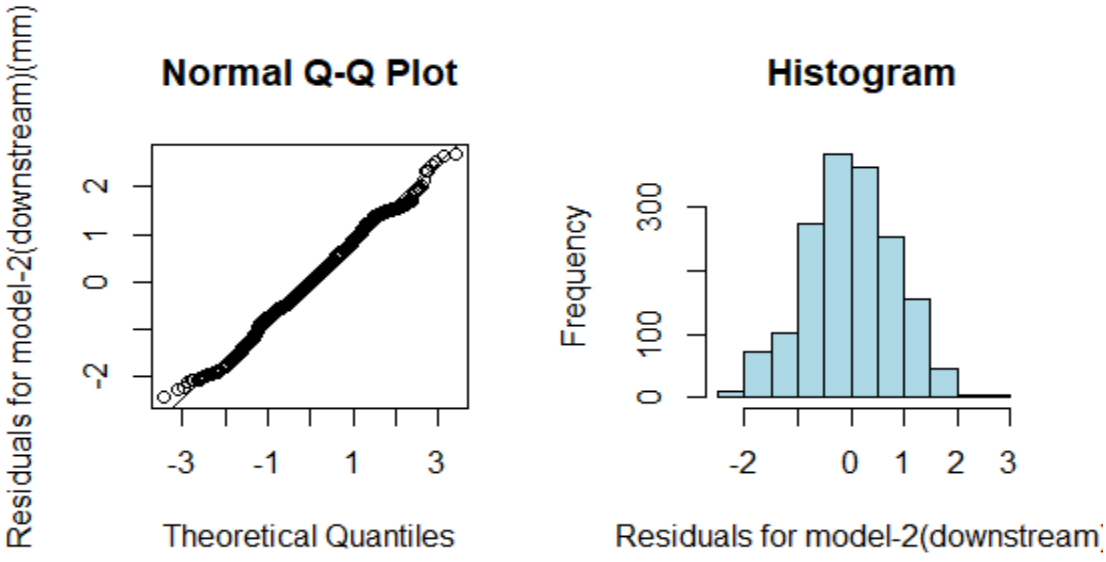

Figure 5.16 - Q-Q plot & histogram of the residuals for the downstream segmented model

The segmented fit improved the normality of the residuals for the downstream model significantly, whereas it marginally reduced the normality of the residuals in the upstream segmented model.

5.4 Conclusions

Overall, the segmented model improved the Adj.R-squared and the Pred R-Squared of both the upstream as well as the downstream models and reduced the overall standard errors by almost 10%. Also, it improved the normality of the residuals for the downstream model. Therefore it can be said that model-2 performs better than model-1. The fit for the downstream model is less accurate than the upstream model because of the higher hysteresis of the piezometer head downstream, and a cubic polynomial can only fit the hysteresis curve approximately. It can be said that the delayed response of the piezometer levels to the reservoir, if not included, leads to more significant residuals and a mediocre fit. Including the delayed response of the piezometer level to the reservoir water level can significantly improve the fit and reduce the autocorrelation.

6 Conclusions and recommendations for future research

6.1 Conclusions

The results indicate that all provide a good fit but the HST_LT model provides the best fit and reduces the residual scattering more than all other proposed models. The segmented model is speculated to show better performance in cases where the hydrostatic component dominates the displacements, but in this study, the seasonal component of the model dominated the regression analysis. Therefore, it did not show much improvement in the fit. The study proves that the HST model is insufficient in cases of extremely cold or hot seasons and that adding lagged variables of air temperature improves the fit of the model significantly. The non-parametric seasonal curve can account for some fluctuations since the curve is computed using seasonal data of all the years in the dataset but it cannot predict the short-term changes accurately.

The results from the piezometer models confirm that the displacements in the dam are an essential predictor of the piezometer level and that the displacements increase sharply after a certain piezometer level. The relationship between the displacements and the piezometer level is non-linear, and a segmented model improves the fit and reduces the overall residuals. The fit for the downstream model is less accurate than the upstream model because of the higher hysteresis of the piezometer head downstream. Therefore, delayed effects of the water level should be introduced in the model to improve its accuracy.

6.2 Recommendations for future use

Since we know that lagged air temperatures help to improve the fit of the HST model, future research should focus on finding out the response function of the dam to the thermal effects or try to approximate it as accurately as possible. Also, since some part of the dam is submerged under water. Therefore, water temperatures at different elevations could help improve the accuracy of the model. Another problem faced by the HST model is that the seasonal components and the hydrostatic components are mostly correlated since the water level in the reservoir is regulated seasonally by most of the dam owners. This results in autocorrelation in the residuals which revokes an assumption made during multi-linear regression. Methods that do not require such assumptions need to be tested to check their applicability.

To model the hysteresis behavior accurately, there is a need for a response function of the dam piezometer level with respect to the increasing and decreasing water levels in the reservoir. Also, the displacements and the water level are correlated with other. Therefore, methods that do not require the assumption of the independence of variables need to be tested.

References

Benvenuto, N., & Piazza, F. (1992). The backpropagation algorithm. IEEE Transactions on Signal Processing, 40(4), 967–969.

Bonelli, S., & Royet, P. (2001). Delayed response analysis of dam monitoring data. In ICOLD European symposiumon dams in a European context. Norway.

Bonelli et al. (2001). Delayed response analysis of temperature effect. In 6th ICOLD Benchmark Workshop on Numerical Analysis of Dams (pp. 1–9). Salzburg, Austria.

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. Journal of the American Statistical Association, 83(403), 596–610.

De Branges, L. (1959). The Stone-Weierstrass Theorem. American Mathematical Society, *10*(5), 822–824.

Dibiagio, E. (2000). Monitoring of dams and their foundations. XXth International congress on large dams. Beijing.

Gomes, A. F. ., & Matos, D. . (1985). Quantitative analysis of dam monitoring results. State of the art, applications and prospects. Proceedings of the 15th International Congress on Large Dams. Lausanne, Switzerland, *1*, 749–761.

Greene, W. W. H. . (2000). Econometric analysis. Prentice Hall (Vol. 97).

ICOLD. (1995). Dam failures - Statistical Analysis. In Bulletin 99 (p. 73pp). Paris.

ICOLD. (1998). Dam-Break Flood Analysis – Review and Recommendations. Bulletin 111.

ICOLD. (2000). Automated Dam Monitoring Systèmes D ' Auscultation Automatique Des Barrages. In International commission on large dams.

International Commission On Large Dams. (2011). Constitution, (January 2002), 1–21.

Jacoby,B.(2005). RegressionIII: Advanced Methods.

Kalkani, E. (1989). Polynomial Regression to Forecast Earth Dam Piezometer Levels. Journal of Irrigation & Drainage Engineering, 115(4), 545–555.

Léger, P., & Leclerc, M. (2007). Model for Concrete Dams. Journal of Engineering Mechanics, 133(3), 267–277.

Mata, J. (2011). Interpretation of concrete dam behaviour with artificial neural network and multiple linear regression models. Engineering Structures, 33(3), 903–910.

Mata et al. (2014). Constructing statistical models for arch dam deformation. Structural Control and Health Monitoring, 21(3), 423–427.

Montgomery, D. C., & Runger, G. C. (2003). Applied Statistics and Probability for Engineers.

National Research Council. (1983). Safety of Existing Dams. Washington D.C.: National Acedemy Press.

Nedushan, B. A. (2002). Multivariate statistical analysis of monitoring data for concrete dams. McGill University.

Penot, I., & Fabre, J. (2009). Analysis and modelling of concrete dams behavior taking into account the air temperature: Method H.S.T Thermal. In 23rd International congress on large dams, Brasil (p. Q.91-R.60). Brasil.

Pipitone, C., Maltese, A., Dardanelli, G., Brutto, M. Lo, & Loggia, G. La. (2018). Monitoring Water Surface and Level of a Reservoir Using Different Remote Sensing Approaches and Comparison with Dam Displacements Evaluated, 1–24.

Platt, I., Hagedorn, M., & Woodhead, I. (2011). The Use of Optical Fibre Sensors in Dam Monitoring, 233–251.

Salazar, F. (2017). A machine learning based methodology for anomaly detection in dam behaviour. PhD Thesis Universitat Politecnica de Catalunya

Salazar, F., Morán, R., Toledo, M., & Oñate, E. (2017). Data-Based Models for the Prediction of Dam Behaviour: A Review and Some Methodological Considerations. Archives of Computational Methods in Engineering, 24(1).

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry, 36(8), 1627–1639.

Tatin, M., Briffaut, M., Dufour, F., Simon, A., & Fabre, J. (2013). Thermal displacements of concrete dams: Finite element and statistical modelling. In 9th ICOLD European club symposium (p. B.24). Venice.

Tatin, M., Briffaut, M., Dufour, F., Simon, A., & Fabre, J. P. (2015). Thermal displacements of concrete dams: Accounting for water temperature in statistical models. Engineering Structures, 91(November), 26–39.

Willm, G., & Beaujoint, N. (1967). Les Methods de Surveillance des Barrages au Service de la Production Hydrauliqe d'Electricite de France, Problemes Anciens et Solution Nouvelles. In IXth International Commission on Large Dams, Q. 34, R. 30, Istanbul (pp. 529–550). Istanbul.

Yu, H., Wu, Z. R., Bao, T. F., & Zhang, L. (2010). Multivariate analysis in dam monitoring data with PCA. Science China Technological Sciences, 53(4), 1088–1097.

Zhao, W. D. (2003). Statistical analysis of monitoring data for Daniel johnson dam. Masters thesis, McGill University.