

Dissertation presented to the Instituto Tecnológico de Aeronáutica, in partial fulfillment of the requirements for the degree of Master of Science in the Graduate Program of Ciência e Tecnologia Espacial, Field of Propulsão Nuclear e Hipersônica.

**Renan Santos Barbosa**

## **ADDITIVE MODELS TO PREDICT CRITICAL HEAT FLUX IN SUBCOOLED FLOW BOILING**

Dissertation approved in its final version by signatories below:

Prof. Dr. Guilherme Borges Ribeiro  
Advisor

Prof<sup>a</sup>. Dr<sup>a</sup>. Camila Pedroso Estevam de Souza  
Co-advisor

Prof. Dr. John von Neumann  
Pro-Rector of Graduate Courses

Campo Montenegro  
São José dos Campos, SP - Brazil  
2024

**Cataloging-in Publication Data**  
**Documentation and Information Division**

Santos Barbosa, Renan

Additive Models to Predict Critical Heat Flux in Subcooled Flow Boiling / Renan Santos  
Barbosa.

São José dos Campos, 2024.  
80f.

Dissertation of Master of Science – Course of Ciência e Tecnologia Espacial. Area of Propulsão Nuclear e Hipersônica – Instituto Tecnológico de Aeronáutica, 2024. Advisor: Prof. Dr. Guilherme Borges Ribeiro. Co-advisor: Profª. Drª. Camila Pedroso Estevam de Souza.

1. Critical Heat FLux. 2. Generalized Additive Models. 3. Prediction. I. Instituto Tecnológico de Aeronáutica. II. Title.

**BIBLIOGRAPHIC REFERENCE**

SANTOS BARBOSA, Renan. **Additive Models to Predict Critical Heat Flux in Subcooled Flow Boiling**. 2024. 80f. Dissertation of Master of Science – Instituto Tecnológico de Aeronáutica, São José dos Campos.

**CESSION OF RIGHTS**

AUTHOR'S NAME: Renan Santos Barbosa

PUBLICATION TITLE: Additive Models to Predict Critical Heat Flux in Subcooled Flow Boiling.

PUBLICATION KIND/YEAR: Dissertation / 2024

It is granted to Instituto Tecnológico de Aeronáutica permission to reproduce copies of this dissertation and to only loan or to sell copies for academic and scientific purposes. The author reserves other publication rights and no part of this dissertation can be reproduced without the authorization of the author.

---

Renan Santos Barbosa  
Rua Rui Barbosa, 146  
14.870-300 – Jaboticabal–SP

# **ADDITIVE MODELS TO PREDICT CRITICAL HEAT FLUX IN SUBCOOLED FLOW BOILING**

**Renan Santos Barbosa**

Thesis Committee Composition:

Prof. Dr.	Alan Turing	Presidente	-	ITA
Prof. Dr.	Guilherme Borges Ribeiro	Advisor	-	ITA
Prof <sup>a</sup> . Dr.	Camila Pedroso Estevam de Souza	Co-advisor	-	UWO
Prof. Dr.	Linus Torwald		-	UXXX
Prof. Dr.	Richard Stallman		-	UYYY
Prof. Dr.	Donald Duck		-	DYSNEY
Prof. Dr.	Mickey Mouse		-	DISNEY

**ITA**

Dedico este trabalho e seus frutos aos meus orientadores, cuja paciência foi fundamental, e à minha esposa, minha fortaleza nos momentos difíceis. Sem eles, não teria alcançado tanto na vida pessoal e profissional.

*"All models are wrong, but some are useful."*  
— GEORGE E.P. BOX

# Resumo

O fluxo crítico de calor (FCC) é o valor no processo de ebulação ou resfriamento no qual o coeficiente da transferência de calor é reduzido drasticamente, ocasionando em um abrupto aumento da temperatura da superfície aquecida/resfriada devido a presença de bolhas ou camadas de vapor ao redor da superfície. Devido as dificuldades em elaborar experimentos e discordâncias sobre métodos de medição e avaliação do fluxo crítico de calor, alguns métodos, como *lookup tables*, correlações utilizando conceitos físicos e modelos de aprendizado de máquina buscam realizar previsões do fluxo crítico de calor de forma otimizada. Modelos aditivos são um tipo de modelo de regressão que utilizam uma estrutura aditiva para definir o tipo de relação existente entre as variáveis de entrada e o FCC para realizar suas previsões. Neste trabalho, diferentes tipos de modelos aditivos serão comparados com métodos consolidados na literatura de previsão do fluxo crítico de calor. Além disto, serão apresentados métodos de interpretação e avaliação dos modelos aditivos ajustados, com o objetivo de demonstrar como os modelos aditivos podem ser utilizados como uma ferramenta alternativa aos modelos teóricos apresentados. Modelos aditivos quantilicos também serão apresentados para o entendimento e previsão de valores extremos do FCC.

# Abstract

The critical heat flux (CHF) is the heat flux value in the boiling or cooling process in which the heat transfer decreases, and the heated surface temperature rises rapidly due to factors such as vapor films and bubbles. Due to the difficulties in elaborate experiments and disagreements about measurement and evaluation techniques related to CHF, some methods, like lookup tables, physical correlations, and machine learning, aim to predict the CHF. Additive models are regression models that use an additive structure to define the relationship between the input variables and the CHF to make predictions. This work will compare multiple additive models with consolidated predictive methods in critical heat flux prediction literature. In addition, methods to evaluate and interpret additive models will be shown to demonstrate that additive models can be used as an alternative tool to predict and infer CHF. Furthermore, quantilic additive models will be shown to study the prediction of extreme values of the CHF.

# List of Figures

FIGURE 1.1 – Flow patterns and heat transfer regions in a heated tube for low and high fluxes velocities of mass. (HALL, 1999) . . . . .	16
FIGURE 2.1 – Subset of Groeneveld’s 2007 LUT (citar) . . . . .	22
FIGURE 2.2 – Example of line formula . . . . .	24
FIGURE 2.3 – Pressure versus CHF in KPa of Zhao’s subcooled data (ZHAO <i>et al.</i> , 2021) . . . . .	24
FIGURE 2.4 – CHF distribution analysis in Zhao’s dataset (ZHAO <i>et al.</i> , 2021) . . .	28
FIGURE 2.5 – GAM and GLM fitted line in Zhao’s dataset (citar) . . . . .	30
FIGURE 3.1 – Distribution of the modeling dataset, on the lower diagonal, a scatterplot of two variables is shown, the diagonal shows the density plot of one variable, and the upper diagonal presents Pearson’s correlation between two variables. . . . .	38
FIGURE 3.2 – Relationship between the pressure and the CHF for each data source, the blue line indicates a default GAM model available by the <i>ggplot2</i> library. . . . .	39
FIGURE 3.3 – Relationship between the mass flux and the CHF for each data source, the blue line indicates a default GAM model available by the <i>ggplot2</i> library. . . . .	39
FIGURE 3.4 – Relationship between the thermodynamic quality and the CHF for each data source, the blue line indicates a default GAM model available by the <i>ggplot2</i> library. . . . .	40
FIGURE 3.5 – Relationship between the heated length and the CHF for each data source, the blue line indicates a default GAM model available by the <i>ggplot2</i> library. . . . .	40

FIGURE 3.6 – Relationship between the diameter and the CHF for each data source, the blue line indicates a default GAM model available by the <i>ggplot2</i> library. . . . .	41
FIGURE 3.7 – Distribution of the CHF for each data source used in this work. . . . .	42
FIGURE 4.1 – Diagram of a simple splitting of the dataset and the $k$ -fold cross-validation method for $k = 10$ , the blue boxes indicate the data used to fit the model, the red ones indicate the data to be used as a test dataset to evaluate the predictive performance of the fitted model. . . . .	44
FIGURE 5.1 – Measured versus predicted CHF values by the models with the ** mark on Table 5.1 . . . . .	50
FIGURE 5.2 – Model’s Relative Error for the measured CHF on the validation dataset. . . . .	51
FIGURE 5.3 – Model’s Relative Error for the measured pressure on the validation dataset. . . . .	51
FIGURE 5.4 – Model’s Relative Error for the measured mass flux on the validation dataset. . . . .	52
FIGURE 5.5 – Model’s Relative Error for the measured thermodynamic quality on the validation dataset. . . . .	52
FIGURE 5.6 – Model’s Relative Error for the measured heated diameter on the validation dataset. . . . .	53
FIGURE 5.7 – Model’s Relative Error for the measured heated length on the validation dataset. . . . .	53
FIGURE 5.8 – One-dimensional smoothing function ( $s("P")$ ) fitted by the best GAM model in predictive performance for the pressure. . . . .	55
FIGURE 5.9 – One-dimensional smoothing function ( $s("G")$ ) fitted by the best GAM model in predictive performance for the mass flux. . . . .	56
FIGURE 5.10 –One-dimensional smoothing function ( $s("X")$ ) fitted by the best GAM model in predictive performance for the thermodynamic quality. . . . .	56
FIGURE 5.11 –One-dimensional smoothing function ( $s("D")$ ) fitted by the best GAM model in predictive performance for the heated diameter. . . . .	57
FIGURE 5.12 –One-dimensional smoothing function ( $s("L")$ ) fitted by the best GAM model in predictive performance for the heated length. . . . .	57

---

FIGURE 5.13 –Two-dimensional tensor functions ( $t()$ ) fitted by the best GAM model in predictive performance. . . . .	59
FIGURE 5.14 –Approximate Predictions Interval for the best GAM 7 model fitted in the validation dataset. The dots represent the measured CHF, and the red line is the predicted CHF. The black lines and the grey area represent the approximate 95% prediction interval. . . . .	61
FIGURE 5.15 –Residual analysis of the best GAM 7 model fitted. . . . .	62
FIGURE 5.16 –One-dimensional smoothing functions for the pressure ( $s("P")$ ) fitted by QGAM model for the quantile $\tau = 0.95$ . . . . .	64
FIGURE 5.17 –One-dimensional smoothing functions for the mass flux ( $s("G")$ ) fitted by QGAM model for the quantile $\tau = 0.95$ . . . . .	64
FIGURE 5.18 –One-dimensional smoothing functions for the thermodynamic quality ( $s("X")$ ) fitted by QGAM model for the quantile $\tau = 0.95$ . . . . .	65
FIGURE 5.19 –One-dimensional smoothing functions for the heated diameter ( $s("D")$ ) fitted by QGAM model for the quantile $\tau = 0.95$ . . . . .	65
FIGURE 5.20 –One-dimensional smoothing functions for the heated length ( $s("L")$ ) fitted by QGAM model for the quantile $\tau = 0.95$ . . . . .	66
FIGURE 5.21 –Two-dimensional tensor functions ( $t()$ ) fitted by the QGAM model for the quantile $\tau = 0.95$ . . . . .	68
FIGURE 5.22 –Predicted $\tau = 0.95$ quantile versus measured CHF. The data was ordered by the lower to the higher value of the CHF to facilitate the visualization. . . . .	70
FIGURE A.1 –Distribution of Groeneveld's (GROENEVELD, 2019) dataset, on the lower diagonal, a scatterplot of two variables is shown, the diagonal shows the density plot of one variable, and the upper diagonal presents Pearson's correlation between two variables. . . . .	78
FIGURE A.2 –Distribution of Zhao's (ZHAO <i>et al.</i> , 2021) dataset, on the lower diagonal, a scatterplot of two variables is shown, the diagonal shows the density plot of one variable, and the upper diagonal presents Pearson's correlation between two variables. . . . .	79
FIGURE A.3 –Distribution of Inasaka's (INASAKA; NARIAI, 1992) dataset, on the lower diagonal, a scatterplot of two variables is shown, the diagonal shows the density plot of one variable, and the upper diagonal presents Pearson's correlation between two variables. . . . .	80

# List of Tables

TABLE 2.1 – Results from the simple linear model fitted . . . . .	27
TABLE 3.1 – Summary of Groeneveld’s (GROENEVELD, 2019) subcooled flow dataset.	35
TABLE 3.2 – Summary of Zhao’s (ZHAO <i>et al.</i> , 2021) subcooled flow dataset . . .	36
TABLE 3.3 – Summary of Inasaka and Nariai (INASAKA; NARIAI, 1992) subcooled flow dataset . . . . .	36
TABLE 3.4 – Summary of the modeling dataset used in this work. . . . .	37
TABLE 4.1 – Additive models evaluated in this work, varying the distribution, link function, relationship structure, and quantile when possible . .	46
TABLE 5.1 – Results of the models fitted in this work. The ** mark indicates the models with lower $RMSE_{val}$ for the four categories of models: lookup table, Hall and Mudawar correlation, GAMs, and QGAMs. The † indicates the influence of the 10th fold in the $RMSE_{cv}$ on the GAM 7 model, which without this fold, the $RMSE_{cv}$ is 32.68. .	49

# List of Abbreviations and Acronyms

CHF	critical heat flux
DNB	departure from nucleate boiling
GLM	generalized linear model
GAM	generalized additive model
LUT	lookup table
LM	linear model
ME	mean error
MAE	mean absolute error
QGAM	quantile generalized additive model
RE	relative error
RMSE	root mean squared error

# List of Symbols

- $x_o$  thermodynamic quality at the output  
 $P$  pressure  
 $G$  mass flux  
 $D$  tube diameter  
 $L$  heated length of the tube  
 $B_o$  boiling number,  $CHF_{measured}/Gh_{fg}$   
 $CHF$  critical heat flux  
 $W_{eD}$  Weber number,  $G^2 D / \rho_f \gamma$   
 $\rho_f$  fluid density at saturation temperature  
 $\rho_g$  fluid saturated vapor  
 $\gamma$  surface tension  
 $y$  observed response variable  
 $Y$  random variable associated to the response variable  
 $\mu$  expected value  
 $\beta$  coefficient of an additive model  
 $\mathbf{x}$  vector of input variables  
 $\sigma^2$  random variable variance  
 $\epsilon$  random error associated with a model  
^ above any symbol indicates the estimator  
– above any symbol indicates the mean observed value  
~ indicates a random variable "follows" a certain distribution  $\mathcal{L}$  likelihood function

# Contents

1	INTRODUCTION . . . . .	16
1.1	Problem Statement . . . . .	16
1.2	Machine Learning in CHF Prediction . . . . .	17
1.3	Objectives . . . . .	18
2	CRITICAL HEAT FLUX PREDICTION . . . . .	20
2.1	Hall-Mudawar Subcooled CHF Correlation . . . . .	21
2.2	Groeneveld's 2006 lookup table (LUT) . . . . .	22
2.3	Additive Models . . . . .	23
2.3.1	Linear Regression . . . . .	23
2.3.2	Generalized Linear Model . . . . .	28
2.3.3	Generalized Additive Models . . . . .	30
2.3.4	Quantile Generalized Additive Models . . . . .	32
3	CHF DATASETS . . . . .	34
3.1	Groeneveld's Dataset . . . . .	34
3.2	Zhao's Dataset . . . . .	35
3.3	Inasaka and Nariai Dataset . . . . .	36
3.4	Modeling Dataset . . . . .	37
4	METHODOLOGY . . . . .	43
4.1	Data Splitting . . . . .	43
4.2	Evaluation Metrics . . . . .	44
4.3	Model fitting . . . . .	45

5	RESULTS . . . . .	48
5.1	Model Comparisons . . . . .	48
5.2	Best GAM Model Interpretation . . . . .	54
5.3	QGAM for Extreme Values . . . . .	62
6	CONCLUSION . . . . .	71
	BIBLIOGRAPHY . . . . .	73
	APPENDIX A – COMPLEMENTARY INFORMATION OF THE USED DATASETS . . . . .	77
	A.1 Distribution Plots of the Used Datasets . . . . .	77

# 1 Introduction

## 1.1 Problem Statement

The critical heat flux can be interpreted as the critical value in which the boiling heat transfer decreases, leading to a fast rise in the heated surface temperature. A thin insulating vapor layer forms on the heated surface at this critical threshold, inhibiting direct contact between the liquid and the surface. As a result, the wall temperature rises rapidly, leading to potential material damage, reduced efficiency, and the possibility of structural failure as in typical water-cooled reactors (GROENEVELD *et al.*, 2007).

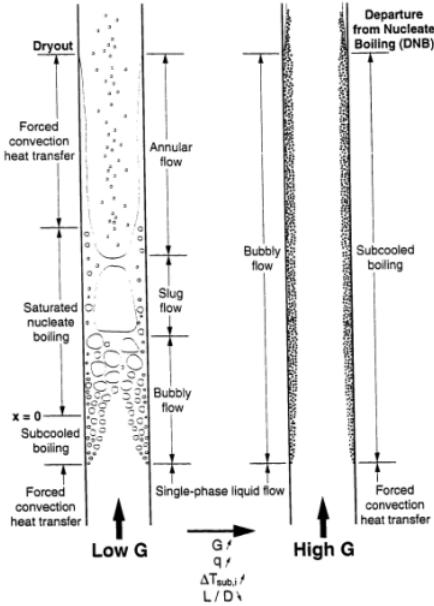


FIGURE 1.1 – Flow patterns and heat transfer regions in a heated tube for low and high fluxes velocities of mass. (HALL, 1999)

Figure 1.1 presented in (HALL, 1999) exemplifies two conditions of CHF occurrence for a uniformly heated tube. The first condition on the left of the figure in the presence

of low mass flux is associated with saturated boiling, in which liquid film dryout is the mechanism related to low values of CHF. On the other hand, the second condition on the right side is related to the subcooled boiling, in which the CHF occurs while the thermodynamic quality is still below zero. This condition is also called departure from nucleate boiling (DNB).

The subcooled boiling is defined by (HALL, 1999) as “the region where the thermodynamic equilibrium quality, averaged over the tube cross-sectional area, is less than zero.” (GROENEVELD, 2019) on the construction of his lookup table considered a subcooled condition in which the is less than zero ( $x_o < 0$ ) in the outlet of the tube. Saturated boiling, on the other hand, is related to positive values of thermodynamic quality.

Despite many studies on the critical heat flux (GROENEVELD, 2019), this topic can still be improved and discussed. An accurate estimate of the CHF is crucial in the design and safety assessment of energy systems reliant on high-temperature processes found in thermal power plants, industrial boilers, and aerospace propulsion systems. Understanding and predicting CHF is essential in these applications to prevent catastrophic failures caused by local overheating and subsequent material damage. By accurately estimating the CHF, engineers can optimize heat transfer mechanisms, determine appropriate cooling strategies, and ensure the safe operation of these systems. Thus, predicting CHF enables the identification of operational limits, guides the design of heat exchange surfaces, and assists in selecting appropriate coolant flows, ultimately leading to enhanced efficiency, reliability, and safety of energy systems operating under high-temperature conditions (GROENEVELD *et al.*, 2007).

The studies developed in (BARBOSA *et al.*, 2023) discussed the potential of using generalized additive models when dealing with the problem of CHF prediction. This work selected a more specific condition of the CHF due to the difficulty of making a model with good predictive performance in a large range of conditions. This condition is related to the observation of the CHF in subcooled flow boiling.

## 1.2 Machine Learning in CHF Prediction

Machine learning models to predict CHF are already being heavily studied. The work from (GROSFILLEY, 2022), which compared Groeneveld’s lookup table (GROENEVELD *et al.*, 2007) with support vector regression, gaussian process regression, and neural networks in which the neural network obtained better predictive results. (KWON; CHANG, 2000) fitted a neural network in the subcooled flow boiling for high CHF prediction problem, comparing its results with mechanistic models and obtaining predictive results favorable to the neural network algorithm.

Zhao (ZHAO *et al.*, 2021) proposed a hybrid approach in which the machine learning model (neural networks and random forests) was fitted to improve the predictions made by a physics-driven model. A similar work to evaluate hybrid models was conducted by (KHALID *et al.*, 2023), which evaluated different machine learning models combined with lookup tables to improve the CHF predictions.

(ZUBAIR *et al.*, 2022),

In addition to the previous works, the reviews from (RASHIDI *et al.*, 2022) and (CHU *et al.*, 2024) explored the use of machine learning in the boiling heat transfer, showing the benefits taken from the use of these techniques not only in the prediction of CHF but in others applications too. (JIANG; LIU, 2018) reviewed the use of machine learning specifically in the CHF prediction. For the three above-cited works, it was noticed that neural networks and support vector regression were the main algorithms applied to the critical heat flux prediction problem.

In his famous paper, Breiman (BREIMAN, 2001) discussed the existence of two cultures in statistical modeling: the data modeling culture, which assumes a model format that the data will be fitted and determines the type of relationship between the inputs and the response variable; and the algorithmic modeling culture which considers the relationship to be complex and unknown, the work in this culture is to find an algorithm that better predicts the output based on a given input. Most of the actual works related to the CHF prediction fell on the second culture, in which the prediction is the main focus. However, most of these algorithmic models, also called black-boxes, have a disadvantage: their lack of interpretability of the results. This work will focus on the first type of modeling, trying to find a method that can obtain good predictions and have tools to understand the fitted model.

### 1.3 Objectives

This work has as the main objective the use of additive models to predict CHF on subcooled flow. An additive model is a statistical regression model in which using the sum as a structure to develop the model creates the possibility of capturing different types of relationships (linear and non-linear) between inputs and the CHF.

The secondary objectives in this work are:

- Show well-established data-based physical correlation models in the literature as a golden standard to compare with the additive models fitted;
- Construct the concept of additive models, starting with the simple linear model, followed by the generalized linear model, then the generalized additive model, and

the complex quantile generalized additive model;

- Measure the predictive performance of the fitted models;
- Show the interpretation tools available in GAMS and QGAMS models and how one can take advantage of this information in the design of an automated system to predict the CHF;
- Discuss the usage of QGAMS to predict extreme conditions of the CHF instead of distribution around the mean like the ones fitted by GAMS.

At the end of this dissertation, all the topics above are expected to be covered, making it a guide for those interested in using additive models to predict CHF.

## 2 Critical Heat Flux Prediction

Methods to predict the critical heat flux date back to the '50s (GROENEVELD, 2019), and the need for more sophisticated methods arose as nuclear reactors worldwide increased. (GROENEVELD *et al.*, 2007) indicated the existence of more than 500 correlations related to the prediction of CHF, (HALL, 1999) compared their proposed correlations with more than 70 specific subcooled correlations. The exhaustive number of CHF correlations can make using them difficult or the choice of which correlation to be used by a researcher can get hard. In this perspective, the studies developed by (GROENEVELD *et al.*, 2007), Hall and Mudawar (HALL, 1999), and Kalimullah (KALIMULLAH *et al.*, 2012) are important to get evidence of the most relevant methods to predict CHF.

From the modeling perspective, CHF prediction methods can be classified into physical, statistical and machine learning, and hybrid models. The physical ones use physical properties and analytical tools to develop CHF correlations. The lookup table from Groeneweld et al. (2007), the proposed model by Caira (CAIRA *et al.*, 1995), and the correlation proposed by Hall and Mudawar (HALL; MUDAWAR, 2000b) are examples of the physical methods. Machine learning tools can be used to predict CHF, like the studies by this author (BARBOSA *et al.*, 2023) related to the use of generalized additive models to predict CHF. Statistical and machine learning methods rely on the use of the trends from data and the relationships defined by the selected predictive method. Hybrid models use the knowledge from physical properties and the prediction capability from learning models combined to get better predictive performance than other methods (ZHAO *et al.*, 2021).

In the following chapters, some of these predictive methods will be shown; in Sections 2.1 and 2.2, physical methods will be explored, and how one can calculate the CHF with them. The other sections, beginning from Section 2.3, will explore the use of additive models (machine learning methods) and their characteristics. It is expected that in the following chapters, an understanding of the methods will enable one to apply them to predict the CHF with accuracy and compare their results.

## 2.1 Hall-Mudawar Subcooled CHF Correlation

Hall and Mudawar (HALL; MUDAWAR, 2000b) proposed two CHF correlations, one based on inlet and the other on outlet conditions for water in a uniformly heated tube. To develop these correlations a subset of the PU-BTPFL dataset (HALL; MUDAWAR, 2000a) containing 4860 data points of subcooled flow boiling data was used. Both correlations assume a non-dimensional form of:

$$Bo = f \left( We_D, \frac{\rho_f}{\rho_g}, x_o \right), \quad (2.1)$$

the outlet conditions CHF correlation is given by Equation (2.2):

$$Bo = C_1 We_D^{C_2} \left( \frac{\rho_f}{\rho_g} \right)^{C_3} \left[ 1 - C_4 \left( \frac{\rho_f}{\rho_g} \right)^{C_5} x_o \right], \quad (2.2)$$

this correlation is valid in the parametric range:  $0.25 \leq D \leq 15.0 \text{ mm}$ ,  $300 \leq G \leq 30,000 \text{ kg m}^{-2} \text{ s}^{-1}$ ,  $1 \leq P \leq 200 \text{ bar}$ , and  $-1.00 \leq x_o \leq -0.05$ .

Using energy balance equations Hall and Mudawar (HALL; MUDAWAR, 2000b) transformed the proposed outlet correlation into an inlet conditions correlation given by:

$$Bo = \frac{C_1 We_D^{C_2} \left( \frac{\rho_f}{\rho_g} \right)^{C_3} \left[ 1 - C_4 \left( \frac{\rho_f}{\rho_g} \right)^{C_5} x_{i*} \right]}{1 + 4C_1 C_4 We_D^{C_2} \left( \frac{\rho_f}{\rho_g} \right)^{C_3+C_5} \left( \frac{L}{D} \right)}, \quad (2.3)$$

this correlation has a valid parametric range of:  $0.25 \leq D \leq 15.0 \text{ mm}$ ,  $300 \leq G \leq 30,000 \text{ kg m}^{-2} \text{ s}^{-1}$ ,  $1 \leq P \leq 200 \text{ bar}$ ,  $-2.00 \leq x_i \leq 0.0$ , and  $-1.00 \leq x_o \leq 0.0$ .

Both correlations have 5 adjustable constants  $C_1$ ,  $C_2$ ,  $C_3$ ,  $C_4$ , and  $C_5$ . To determine the optimal values of the constants, nonlinear regression was used and obtained the following values:  $C_1 = 0.0722$ ,  $C_2 = -0.312$ ,  $C_3 = -0.644$ ,  $C_4 = 0.900$ , and  $C_5 = 0.724$ . The optimal values suggested by the authors can be used in both proposed correlations. In their study, Hall and Mudawar (HALL, 1999) compared their correlation with 78 correlations available in the literature at that moment; their proposed inlet conditions correlation was the one that obtained the best results.

In this work, we will compare our fitted models with the correlations of Hall and Mudawar. The fitted additive models are expected to achieve predictive results closer to or better than the above correlations.

## 2.2 Groeneveld's 2006 lookup table (LUT)

Groeneveld's LUT (GROENEVELD *et al.*, 2007) is one of the main methods to predict CHF. The LUT is a normalized dataset to predict CHF on 8mm vertical round tubes. The 2006 LUT is an improved version of the previous 1995 LUT (GROENEVELD *et al.*, 1996), in which new datasets were included and a review of the methods to develop the new LUT.

This method covers the following ranges of parameters:  $0.1 \leq P \leq 20 \text{ MPa}$ ,  $0 \leq G \leq 7500 \text{ kg m}^{-2} \text{ s}^{-1}$ , and  $-0.5 \leq x_o \leq 1$ . The LUT is a discrete value table in which the values of pressure, mass flux, and exit equilibrium quality are necessary to obtain the desired CHF; a subset of the table is shown in Figure 2.1. The first challenge in this project was accessing the table, which was not completely available in the article (GROENEVELD *et al.*, 2007). Only in (GROENEVELD, 2019) was the full table available, but in a format that cannot be used to make data analysis. The data transcription was made and will be available on the *Github*'s project repository of this work.

P kPa	G kg/m <sup>2</sup> /s	X												
		-0.50	-0.40	-0.30	-0.20	-0.15	-0.10	-0.05	0.00	0.05	0.10	0.15	0.20	0.25
100	0	8111	7252	6302	4802	4086	3057	1990	1142	637	415	284	223	188
100	50	8317	7271	6326	5035	4236	3453	2420	1570	1011	784	641	587	553
100	100	8390	7295	6371	5322	4586	3640	2942	2103	1558	1275	1013	885	847
100	300	10698	9288	7795	6020	5009	3865	3196	2479	1961	1707	1317	1177	1172
100	500	12882	10946	9224	6791	5348	3938	3369	2685	2087	1808	1412	1347	1311
100	750	16982	14405	11641	7496	5662	4234	3471	2780	2229	1970	1649	1606	1591
100	1000	19441	16278	13255	8232	5971	4495	3533	3012	2653	2349	2070	2000	1980
100	1500	22781	19225	15465	9100	6603	5358	3741	3524	3166	2917	2635	2572	2467
100	2000	25268	21321	17143	9141	7059	6036	4074	3855	3556	3402	3167	2986	2720
100	2500	28026	23599	18346	9503	7506	6516	4502	4047	3852	3599	3228	3019	2676
100	3000	30294	25465	19383	9779	8063	7088	4826	4182	3976	3389	2968	2706	2369

FIGURE 2.1 – Subset of Groeneveld's 2007 LUT (citar).

As the table has only a discrete range of values, it is necessary to make a linear interpolation between the variables and the CHF from the LUT to obtain the CHF. For tubes with diameters with different values than 8 mm, it is necessary to do a correction on the predicted CHF, done by 2.4:

$$CHF = CHF_{8mm} \left( \frac{D}{0.008} \right)^{-n}. \quad (2.4)$$

In their study, (GROENEVELD *et al.*, 2007) proposed the value of  $n = 0.5$ , otherwise, (TANASE *et al.*, 2009) recommended a new value of  $n = 0.312$  that is close to the study of (CELATA *et al.*, 1996) stated in (KALIMULLAH *et al.*, 2012).

In this work, both values of  $n$  will be evaluated and compared with the proposed fitted models. It is expected that the fitted models achieve predictive results closer to or better

than the LUT since the LUT is a generic method developed for a wide range of conditions of CHF, resulting in weaker results for a specific use case, the subcooled flow boiling CHF.

## 2.3 Additive Models

In a prediction problem when a response variable (CHF) denoted by  $Y \in \mathbb{R}$  and a vector of features (pressure, mass flux, equilibrium quality, ...) denoted by  $\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$  this problem can be classified as a regression problem. A regression model tries to find the relationship between  $Y$  and  $\mathbf{x}$  with the given data. This relationship can be assumed to be linear, non-linear, or a mixture of both, and the approach chosen to find this association determines the type of relationship. Some examples of these techniques are:

- Linear Regression;
- Generalized Linear Models;
- Generalized Additive Models;
- Quantile Generalized Additive Models;

These are called additive models because they use a structure that assumes a relationship given by a sum. In the following subsections, the models are described by order of complexity, starting with the simple linear regression model, and ending with the more complex version, the quantile generalized additive model (QGAM). From a modeling perspective, starting with a simple technique can help better understand the problem to be solved and avoid using a complex method that can achieve results closer to a simpler one (principle of parsimony in modeling).

### 2.3.1 Linear Regression

The formula to draw a line between a variable  $y$  and  $x$  can be given by the following formula:

$$y = a + bx, \quad (2.5)$$

in which the  $a$  value named the “intercept” of the line is the calculated value of  $y$  when  $x = 0$ , the value  $b$  is called the “slope” that indicates the “inclination” of this line. Figure 2.2 shows an example of the Equation 2.5 when  $a = 0.4061$  and  $b = 4.2667$ .

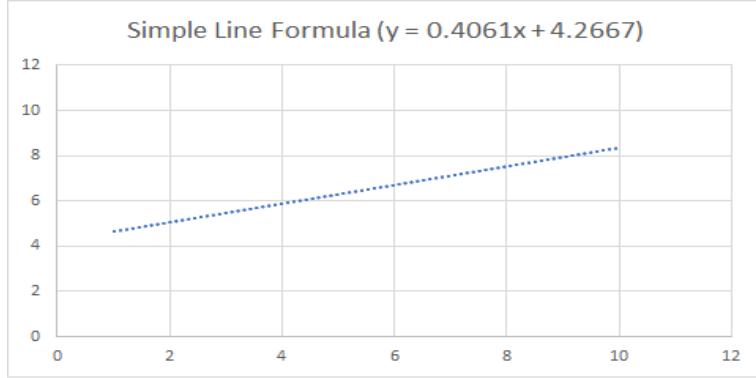
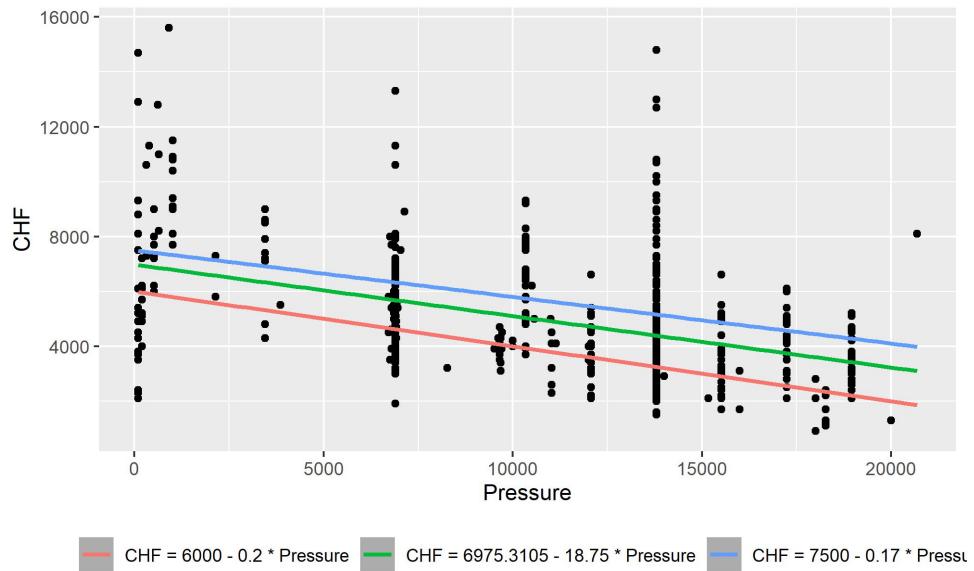


FIGURE 2.2 – Example of line formula

One would like to study the relationship between variables when analyzing collected data. For example, Figure 2.3 shows a scatter plot of the pressure ( $x$ -axis) and the CHF ( $y$ -axis) of Zhao's subcooled CHF data (ZHAO *et al.*, 2021). When observing the data points, a simple line can be drawn to explain the relationship between the variables. However, as shown in Figure 2.3, the number of lines that can be formulated and look feasible can be infinite.

FIGURE 2.3 – Pressure versus CHF in KPa of Zhao's subcooled data (ZHAO *et al.*, 2021)

A statistical method to obtain a “good” estimate of a line equation that represents the relationship between the variables using the collected data can be obtained with the simple linear regression model (MONTGOMERY *et al.*, 2012), which is formulated as follows:

$$Y_i = \mu_i + \epsilon_i, \text{ where } \mu_i = \beta_0 + x_i \beta_1 , \quad (2.6)$$

in which  $i = 1, 2, 3, \dots, n$  is the  $i$ -observation of the data,  $Y_i$  is the random variable associated with the observed response variable  $y_i$ ,  $\mu_i$  is the expected value of the random

variable, i.e.,  $\mu_i = \mathbb{E}(Y_i)$ . In addition,  $\beta_0$  and  $\beta_1$  are unknown coefficients to be estimated, and  $\epsilon_i$  are independent random errors with zero mean and equal variance  $\sigma^2$ .

Two methods that can be used to estimate the values of  $\beta_0$  and  $\beta_1$  are the least squares estimation and the maximum likelihood estimation. In both cases, the estimated values of the  $\beta$ 's are equal, but the assumptions made and the properties available are different. In the least squares method, a measurement of how “good” the estimated  $\beta$ 's must be used. The residual sum of squares (WOOD, 2017; IZBICKI; SANTOS, 2020) is one of these metrics, given by,

$$S = \sum_{i=1}^n (y_i - \mu_i)^2 = \sum_{i=1}^n (y_i - \beta_0 + x_i \beta_1)^2 , \quad (2.7)$$

by minimizing the values of  $S$  in relation to the  $\beta$ 's, it is possible to obtain the estimates of the  $\beta$ 's that minimize the function  $S$ . These estimators  $\hat{\beta}$  of the  $\beta$ 's are:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} , \text{ and} \quad (2.8)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} , \quad (2.9)$$

where  $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ ,  $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$ ,  $S_{xy} = \sum_{i=1}^n (x_i - \bar{x}) y_i$ , and  $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ .

Now, it is possible to obtain the least squares estimates of the  $\beta$ 's using the collected data, and applying Equation 2.10, it is possible to obtain the estimates of each  $\hat{y}_i$ .

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i . \quad (2.10)$$

Using the data shown in Figure 2.3, the linear equation to estimate the CHF is available in Equation 2.11.

$$\widehat{CHF} = 6975.3103 - 0.1875 \times Pressure . \quad (2.11)$$

Although the least squares method can give a way to make predictions of CHF given an input, one may also be interested in evaluating if the estimated  $\beta$ 's are statistically significant (can be considered different from zero), the confidence interval of the  $\beta$ 's and the predicted values can also be estimated to understand better and make interpretations about the fitted model.

To make these tests and analyses, it is necessary to add an assumption in the errors  $\epsilon_i$ , that they are independent and identically distributed (i.i.d.), following a normal prob-

ability distribution, i.e.,  $\epsilon_i \sim N(0, \sigma^2) \Rightarrow Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ . When making these assumptions, the maximum likelihood estimation can be used, which uses the concept of the likelihood function, which is the joint probability distribution as a function of its parameters (CASELLA; BERGER, 2002).

The likelihood function for  $n$  i.i.d. samples of a random variable  $Y$ ,  $\mathbf{y} = y_1, y_2, \dots, y_n$  and unknown parameters  $\beta_0$  and  $\beta_1$  is defined as

$$\mathcal{L}_i(\beta_0, \beta_1; \mathbf{y}) = \prod_{i=1}^n f_{y_i}(y_i; \beta_0, \beta_1). \quad (2.12)$$

In the linear regression problem, it is assumed that the distribution of  $Y$  is i.i.d. following a normal distribution. In this case, the likelihood function can be written as

$$\prod_{i=1}^n f_{y_i}(y_i; \beta_0, \beta_1) = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{y_i - (\beta_0 + \beta_1 x_i)}{2\sigma^2} \right\}. \quad (2.13)$$

The maximum likelihood estimation uses the Equation 2.13 to obtain the estimators of the  $\beta$ 's by maximizing this function with respect to each parameter, usually the log of the function is used since it is a simpler function to be used and its equivalent of the likelihood itself when used to obtain estimators. Another benefit is that this equation can be used analytically when the distribution is well-defined (like in the linear regression case) and numerically when a closed-form equation cannot obtain the estimates.

In the linear regression model, the maximum likelihood estimators of the  $\beta$ 's are the same as the least squares estimation method shown in Equation 2.8 obtaining the same line equation of Equation 2.11. Furthermore, adding the assumption of normality in the errors can make possible the use of hypothesis tests for the  $\beta$ 's and confidence intervals in our predictions. For example, one may be interested in testing if a  $\beta_1 \neq 0$ , i.e., to test if the input variable is statistically significant to influence the value of  $y$ .

Table 2.1 shows the results obtained by the model mentioned before. The first value is the obtained  $\beta$ 's. Next to the coefficients, the "\*" symbol indicates the significance level of the fitted  $\beta$ 's, i.e., if one decides that, at a significance level of  $\alpha = 0.05$ , the variables used are statistically significant to the model and their  $\beta$ 's are  $\neq 0$ . The 95% confidence interval for each  $\beta$  is shown below its coefficient value.

However, adding a distribution in the errors requires evaluating its assumptions to determine if the model fitted is reliable enough and can be used. It is necessary to check the normality assumption in the fitted errors using quantile-quantile plots and normality tests on the errors (MONTGOMERY *et al.*, 2012). A deeper analysis of the fitted residuals (errors) is needed to check independence.

TABLE 2.1 – Results from the simple linear model fitted

<i>Dependent variable:</i>	
	CHF
Pressure	-0.187*** (-0.215, -0.160)
Intercept	6,975.310*** (6,641.389, 7,309.232)

*Note:* \*p-value<0.1; \*\*p-value<0.05; \*\*\*p-value<0.01

Furthermore, the CHF prediction problem uses more than just one variable. In the case of multiple inputs given by the vector  $\mathbf{x} = (x_1, x_2, \dots, x_d)$ , the  $(n, d+1)$  design matrix of  $n$  observed data,  $X$ , is:

$$X = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,d} \\ 1 & x_{2,1} & \dots & x_{2,d} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,d} \end{bmatrix}_{(n,d+1)}, \quad (2.14)$$

in which the first column of ones is used when the fitted model assumes the presence of an intercept ( $\beta_0$ ).

In the multiple inputs case, the vector of coefficients estimates  $\hat{\beta}$  is given by the matrix-vector computation,

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \quad (2.15)$$

the assumptions made from the simple linear regression model can be extended for the case in which multiple inputs are considered.

The methods and concepts briefly described in this section are the foundations for the next ones that will be shown. A new method is developed by changing some assumptions of the linear regression model, either on the distribution of the errors or going beyond linearity. These changes make it possible to fit different types of relationships. The next section will present the generalized linear model, which increases the possibility of new models being fitted.

### 2.3.2 Generalized Linear Model

When analyzing the distribution of the CHF presented in Figure 2.4, the assumption that CHF follows a normal probability distribution does not hold.

Although this assumption is not true for Zhao's dataset, methods to transform the response variable, like the log transformation or the Box-Cox transformation (BOX; COX, 2018) are available. However, direct transformations in the values of the response variable may indirectly change the interpretation of the results and the distribution of the errors when turning the predictions to the original scale.

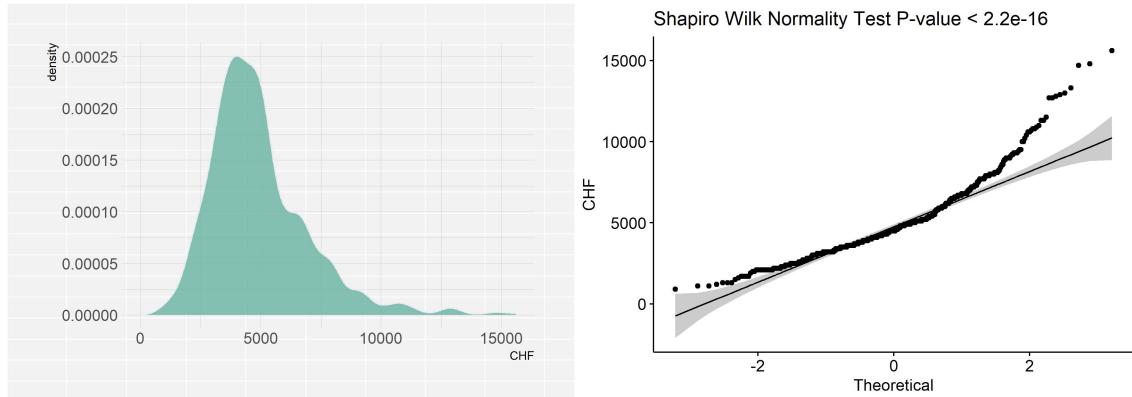


FIGURE 2.4 – CHF distribution analysis in Zhao's dataset (ZHAO *et al.*, 2021)

(NELDER; WEDDERBURN, 1972) proposed a new method called generalized linear models that can be used by defining some components of a regression problem. In the class notes of (DEMETRIO, 2002), it is stated that these components are the following:

1. The random component in the model. Represented by a vector of random variables  $Y_1, Y_2, \dots, Y_n$ . These variables must have the same distribution and be part of the exponential family of distributions;
2. A systematic component determined by the experiment and the desired design of the model to be fitted;
3. A link function  $g(\cdot)$ , that will link the random and the systematic components.

A random variable  $Y_i$  belongs to the exponential family of distributions (MCCULLAGH; NELDER, 1989) if its density can be written as :

$$f(y_i; \theta_i) = \exp\left\{ \frac{1}{a_i(\phi)} [y_i \theta_i - b(\theta_i)] + c(y_i; \phi) \right\}, \quad (2.16)$$

in which  $\phi > 0$  is a known scale parameter,  $a_i(\cdot)$ ,  $b(\cdot)$ , and  $c(\cdot)$  are arbitrary functions,  $\theta_i$

is called the 'canonical' parameter. By theoretical properties of the exponential family, it is possible to determine a generic form of the mean and variance for  $Y_i$ ,

$$\mathbb{E}(Y_i) = \mu_i = b'(\theta_i) , \quad (2.17)$$

$$Var(Y_i) = a_i(\phi)b''(\theta_i) , \quad (2.18)$$

usually  $a_i(\phi) = \phi/w_i$ , where  $w_i$  is a known constant (WOOD, 2017), a list of distributions and its properties are available in (WOOD, 2017, p. 104).

(DEMETRIO, 2002), stated that the systematic component can be written as a linear sum of its effects, i.e.,

$$\eta = X\beta , \quad (2.19)$$

in which  $X$  is the design matrix shown in 2.14,  $\beta$  is the  $d + 1$  vector of the model parameters, and  $\eta$  is called the linear predictor of the model.

The last component to define a GLM is the link function, which can be interpreted as a function to relate the mean with the linear predictor, i.e.,

$$\eta_i = g(\mu_i) , \quad (2.20)$$

in which  $g(.)$  is a differentiable monotonic function. When the chosen  $g(.)$  has the form of  $g(\mu_i) = \theta_i$ , this link function is called the "canonical" link, enabling better theoretical properties of the estimates, easier estimation of the  $\beta$ 's, and interpretation of the parameters (DEMETRIO, 2002, p. 27).

With all these three components determined, it is possible to start using GLMs. For example, using the normal distribution for  $Y$ , a linear predictor as in 2.19, and an identity link function, i.e.,  $\eta = \mu$ , the GLM is transformed into a simple linear regression model. However, as Zhao's data does not follow a normal distribution, the GLM enables using other distributions, like the Gamma or the log-normal distributions, that assume only positive values in the response and can be a possible alternative to the normal distribution.

Unlike linear regression, GLMs cannot be analyzed analytically in all cases. By defining that  $Y$  is from the exponential family, estimating the parameters needs numerical algorithms. Some methods available to estimate the parameters are the Newton-Rapshon (DEMETRIO, 2002, pp. 28-33) and the iteratively re-weighted least squares (IRLS) algorithm (WOOD, 2017, pp. 105-107).

However, it is possible that the explainable variables do not have a linear relationship with the expected value of the CHF. For example, in Figure 2.5, it is possible to observe

that although the linear model can be a good option, indicating a negative relation with the CHF, this negative trend is not regular in some specific pressure regions. To address the case in which the relationship between the inputs and the CHF can be nonlinear, the generalized additive model (GAM) is an alternative. Section 2.3.3 will discuss the additions in a GLM to fit these nonlinear trends with GAMs.

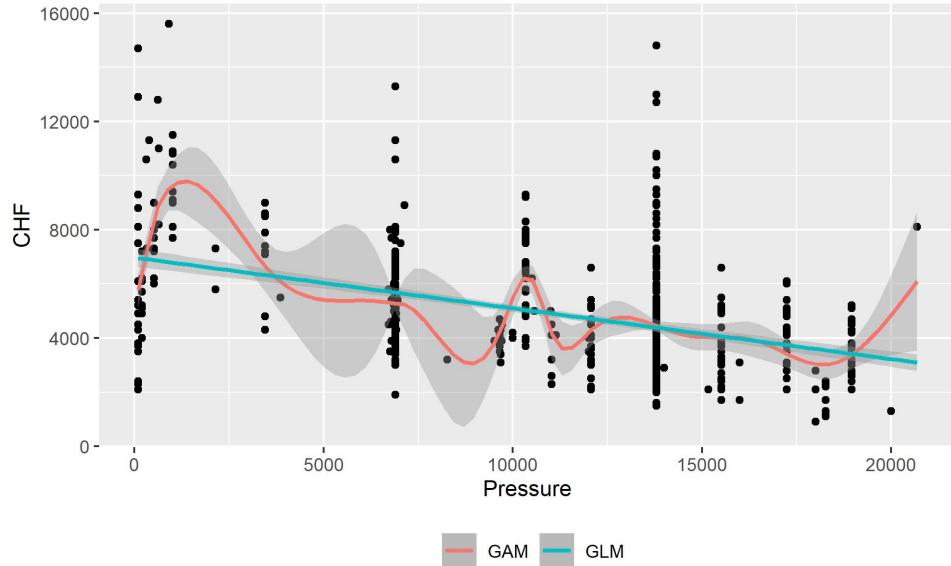


FIGURE 2.5 – GAM and GLM fitted line in Zhao’s dataset (citar)

### 2.3.3 Generalized Additive Models

Introduced by (HASTIE; TIBSHIRANI, 1986), generalized additive models (GAMS) were proposed as an alternative to GLMS, which enables the possibility to fit nonlinear relationships between the desired response and the inputs without an explicit specification of this relation.

A GAM can be defined “as an extension of the generalized linear model (GLM), where in addition to the GLM common characteristics, a sum of smooth functions is added to the model linear predictor, making possible the use of nonlinear relationships between the response variable and its predictors without a parametric specification of these relations” (BARBOSA *et al.*, 2023)

(WOOD, 2017) defined the general structure of a GAM as:

$$g(\mu_i) = A_i\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \dots , \quad (2.21)$$

in which  $A_i\theta$  are terms related to a specific parametric model from a GLM, the other assumptions and terms discussed in the section 2.3.2, like link functions, exponential

family, are still present in the GAMS models.

The main addition made in the GAMS is the term  $f_i$ , which stands for the smooth functions of the covariates. These smooth functions can be univariate and multivariate, and the method to estimate each can differ. (WOOD, 2017) states that this addition makes it possible for a more flexible model to be fitted. However, at a cost, it is necessary now to define how to represent these smoothing functions and their level of smoothness.

One simple way to represent a smoothing function is using *basis* functions, i.e.,

$$f_i(x) = \sum_{j=1}^k b_j(x)\beta_j , \quad (2.22)$$

in which  $b_j$  is a known base function,  $\beta_j$  are the unknown basis parameters to be estimated, and  $k$  is the number of the basis functions.

Substituting Equation 2.22 in 2.21, it can be noted that the model can now be interpreted as a GLM. These basis functions can assume any form, such as polynomials, piecewise linear basis, and cubic regression splines.

However, using basis functions to define a smooth function requires determining the best value of  $k$  that fits the data. One way to avoid this decision is to fix a value of  $k$  high enough to capture the data variability and add a ‘wigginess’ penalty to determine the smoothness of  $f$ .

In general, this penalty term can be written as

$$\sum_i \lambda_i \boldsymbol{\beta}^T \mathbf{S}_i \boldsymbol{\beta} , \quad (2.23)$$

$\lambda_i$  is a smoothing parameter for each  $f_i$ , and  $\mathbf{S}_i$  is a penalty matrix determined by the basis function chosen. The  $\lambda_i$  on Equation 2.23 is written in a way that high values of  $\lambda \rightarrow \infty$  lead to a straight line of the  $f$  function, and  $\lambda = 0$  results in a wiggler  $f$  smoothing function.

By defining the model and the penalty term, we want to find the  $\beta_s$  that maximizes the Equation 2.24 (WOOD, 2017).

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2\phi} \sum_i \lambda_i \boldsymbol{\beta}^T \mathbf{S}_i \boldsymbol{\beta} . \quad (2.24)$$

Given the definitions, when the  $\lambda$ 's are fixed, the parameters  $\beta$ 's can be estimated with penalized iteratively re-weighted least squares (PIRLS) (WOOD, 2017, pp. 148-149). The more complex process in a GAM model is choosing the  $\lambda$ 's.

Two alternatives for choosing the  $\lambda$ 's available are methods based on error measurement like cross-validation, in which part of the data is fitted with a fixed  $\lambda$ , and the remainder is used to evaluate the model's predictive performance; the same model assessed at different levels of  $\lambda$  are then used to get the best value to be used. The other methods are based on Bayesian/mixed models via marginal likelihood maximization or REML (WOOD, 2017).

### 2.3.4 Quantile Generalized Additive Models

The methods above are based on estimating the expected value of  $Y$ ,  $\mu = E(Y)$  and propose the use of a parametric probability distribution to estimate its parameters. However, a distribution cannot define the data collected in many real-world applications. Also, the assumption of homoscedasticity cannot be verified in some cases.

Quantile regression models, on the other hand, aim to fit the quantile  $\tau$  of the response variable  $Y$ ; for example, considering  $\tau = 0.5$ , the fitted model will predict the median value of  $Y$  given a set of inputs.

A more formal definition of the quantilic regression model can be done, given the conditional cumulative function (c.d.f.) formulated as  $F(y|\mathbf{x}) = P(Y \leq \mathbf{x})$ , then the conditional quantile  $q$  is defined as  $q = \inf\{y : F_y(y|\mathbf{x}) \geq \tau\}$ , where  $\tau \in (0, 1)$ . (FASIOLO *et al.*, 2021a) states that the conditional quantile  $q$  can be defined as the minimizer of the following expected loss

$$L(q|\mathbf{x}) = \mathbb{E}(\rho_\tau(y - q)|\mathbf{x}) = \int \rho_\tau(y - q)dF(y|\mathbf{x}) , \quad (2.25)$$

w.r.t.  $q = q(\mathbf{x})$ , where

$$\rho_\tau(z) = (\tau - 1)z\mathbb{1}(z < 0) + \tau z\mathbb{1}(z \geq 0) , \quad (2.26)$$

is the pinball loss.

When a set of  $n$  observed data is collected,  $dF(y)$  can be approximated to its empirical version  $dF_n(y)$ , and the estimator of the quantilic model can be defined as

$$\hat{q} = \operatorname{argmin}_q \frac{1}{n} \sum_{i=1}^n \rho_\tau\{y_i - q(\mathbf{x}_i)\}. \quad (2.27)$$

(FASIOLO *et al.*, 2021a) proposes the use of an additive structure for  $q(\mathbf{x})$  similar to the ones used in the GAMS models, i.e.,  $q(\mathbf{x}) = \sum_{j=1}^m f_j(\mathbf{x})$ .

In addition to the additive structure used in (FASIOLO *et al.*, 2021a) work, their work shows how the pinball loss shown in Equation 2.26 can be suboptimal compared to a generalized version of the loss. Furthermore, (FASIOLO *et al.*, 2021b) provides a simple tool in the R programming language to fit the proposed model in (FASIOLO *et al.*, 2021a) that works similarly to the one available in the *mgcv* package.

In this work, we will fit the QGAM models for multiple values of  $\tau$  that will be 0.05, 0.5, and 0.95. The quantile  $\tau = 0.5$  will be fitted to be compared with the models of the previous chapters. The other quantiles will be fitted to better understand the input conditions related to extreme values of the CHF.

# 3 CHF Datasets

The collection of CHF datasets was one of the main challenges in this project; besides being cited and evaluated, many datasets were not publicly available to be collected and used in this work. The datasets available in (GROENEVELD, 2019; ZHAO *et al.*, 2021) made most of this project possible. However, datasets like (HALL; MUDAWAR, 2000a) that comprise a good amount of subcooled CHF data were unavailable. Different collections of datasets like the ones in (KNOEBEL *et al.*, 1973; CELATA *et al.*, 1993; VANDERVORT *et al.*, 1994) were already cleaned and studied by (GROENEVELD, 2019) and are available in their paper.

This chapter will show and discuss the datasets used in this work and their range of values and unique characteristics. Appendix A shows the distribution of the variables for each dataset and the final dataset used to fit the models and correlations discussed in Chapter 2. The collected datasets use water as the working fluid, and only data associated with the subcooled flow boiling process were used.

## 3.1 Groeneveld's Dataset

Groeneveld made available the dataset to develop the 2006 LUT (GROENEVELD, 2019). The provided dataset contains 25,000 observations and eight variables from the collection of studies available in the literature that were selected and cleaned. The selected data points related to subcooled flow boiling heat flux ( $x_o < 0$ ) and the removal of duplicated data reduced the number of observations to 1,886. Table 3.1 summarizes the variables of Groeneveld's data used in this work.

TABLE 3.1 – Summary of Groeneveld’s (GROENEVELD, 2019) subcooled flow dataset.

Variable	Unit	Min	Max	Mean	Std
$P$	$kPa$	100	20,000	10,664	6,526.20
$G$	$kg\ m^{-2}\ s^{-1}$	345	7,964	3,008	1,648.58
$x_o$	-	-0.497	-0.0003	-0.1039	0.10
$D$	$mm$	2	15.82	8.339	2.07
$L$	$mm$	50	4000	843	646.90
$CHF$	$kWm^{-2}$	643	16,339	5,187	2,743.34

In addition to the summary of the variables, the analysis of the distribution of the variables is shown in Figure A.1. Besides being a database with a good amount of data compared to other studies related to subcooled CHF, this dataset notoriously has some gaps that must be filled.

- The outlet quality  $x_o$  ranges from -0.5 to 0, limiting a fitted model to this range of values of quality. Compared to Groeneveld’s dataset, Hall and Mudawar PU-BTPFL databank (HALL; MUDAWAR, 2000a) their subcooled flow CHF data contains outlet qualities ranging from -2.25 to 0.0 and a CHF correlation (HALL, 1999) with  $x_o$  starting from -1.0 to 0.0 (or -0.05 for the output conditions correlation);
- A similar condition to the vapor quality can be observed with the mass flux variable  $G$ . Groeneveld’s data set for subcooled flow values does not have values of CHF for high fluxes;
- The distribution of  $D$ ,  $L$ , and  $P$  shows more than one mode and valley; this behavior indicates the concentration of the experiments in one specific region of values, and it can make our model get poor predictions in the areas of low density.

## 3.2 Zhao’s Dataset

(ZHAO *et al.*, 2021) made available a dataset containing 1865 observations related to the event called a departure from nucleate boiling (DNB), which occurs in studies of CHF for nuclear power plants. When filtering the dataset for subcooled flow ( $x_o < 0$ ) and the removal of duplicated data, the dataset is reduced to 700 observations.

One addition to this dataset compared to Groeneveld’s is the presence of geometries different from round tubes (588 observations), with data points of the annulus (64 observations) and plate (48 observations) geometries. Other geometries can be seen as an

additional difficulty to the tested models and their predictive capabilities, as they are usually fitted for round tube problems. Table 3.2 and Figure A.2 show a descriptive analysis of the subcooled flow in Zhao’s dataset.

This dataset has characteristics similar to those of Groeneveld’s, with subcooling values concentrated between -0.2 and 0, a low concentration of data on high fluxes, and the presence of modes and valleys of diameter and heated length.

TABLE 3.2 – Summary of Zhao’s (ZHAO *et al.*, 2021) subcooled flow dataset

Variable	Unit	Min	Max	Mean	Std
$P$	$kPa$	100	20,680	10,854	4,861.85
$G$	$kg\ m^{-2}\ s^{-1}$	0	7,975	3,327	1,901.16
$x_o$	-	-0.8667	-0.0007	-0.0912	0.09
$D$	$mm$	1.1	120	17.71	29.30
$L$	$mm$	10	2,591	610.8	533.81
$CHF$	$kWm^{-2}$	900	15,600	4,955	2,136.58

### 3.3 Inasaka and Nariai Dataset

(INASAKA; NARIAI, 1992) conducted experiments on subcooled flow with water and how it affects the CHF. The available dataset contains 30 observations added to other studies from different sources, and their values were used as input to compare different correlations. Table 3.3 summarizes the 30 observation dataset informed by Inasaka and Nairai. Figure A.3 shows the distribution of the collected variables.

TABLE 3.3 – Summary of Inasaka and Nariai (INASAKA; NARIAI, 1992) subcooled flow dataset

Variable	Unit	Min	Max	Mean	Std
$P$	$kPa$	290	1,050	629	257.15
$G$	$kg\ m^{-2}\ s^{-1}$	4,300	30,200	15,230	7,969.604
$x_o$	-	-0.199	-0.051	-0.1232	0.041
$D$	$mm$	3	3	3	0
$L$	$mm$	100	100	100	0
$CHF$	$kWm^{-2}$	7,300	48,000	22,320	9,933.69

### 3.4 Modeling Dataset

The modeling dataset used in this work combines the aforementioned data sources with a total of 2513 observations and six variables, the response variable (CHF), and five explanatory variables: diameter, heated length, mass flux, pressure, and thermodynamic quality. This dataset had possible duplicated values removed since Groeneveld's data comes from multiple sources. Table 3.4 contains the main statistics of the collected dataset, and Figure 3.1 shows the distribution of the variables.

One of the main characteristics of the data is its high variance and sparsity due to the studies of CHF, in which some conditions need to be fixed, like the diameter and the heated length to measure the CHF, making it hard to obtain a range of values of these variables. Usually, one may treat these variables as fixed levels; however, as this work wants to evaluate and predict the CHF in different values of diameter and heated length, this project will treat them as continuous variables that can be assessed at any value.

TABLE 3.4 – Summary of the modeling dataset used in this work.

Variable	Unit	Min	Max	Mean	Std
$P$	$kPa$	100	20,680	10,618	6,223.73
$G$	$kg\ m^{-2}\ s^{-1}$	0	30,200	3,212	2,273.04
$x_o$	-	-0.8667	-0.0003	-0.1015	0.099
$D$	$mm$	1.10	120	10.67	15.45
$L$	$mm$	10	4,000	780.2	631.226
$CHF$	$kWm^{-2}$	643	48,000	5,306	3,282.805

In Figures 3.2,3.3, 3.4, 3.5, and 3.6, the relationship between the explanatory variables and the CHF for each source of data discussed before is shown. It is possible to observe that the pressure variable has a negative relationship with the CHF; however, for low-pressure values, this relationship appears to be positive with a turning point of pressure close to 3000  $kPa$ . Furthermore, a positive relationship can be observed for the mass flux. On the other hand, the thermodynamic quality and the heated length indicate a negative relationship with fluctuations in some areas. A clear relationship between the diameter and the CHF can be seen only in Groeneveld's data source, where a negative relation can be observed.

On all the variables, it is possible to observe that multiple values of the CHF are observed for the same input level value. This behavior can make the fit of a model other than the QGAM less efficient, as the supposition of homoscedasticity in the variance cannot be observed. On the other hand, the QGAM model can take advantage of this behavior by the fit of multiple models on different levels of quantiles; in Chapter 5, the

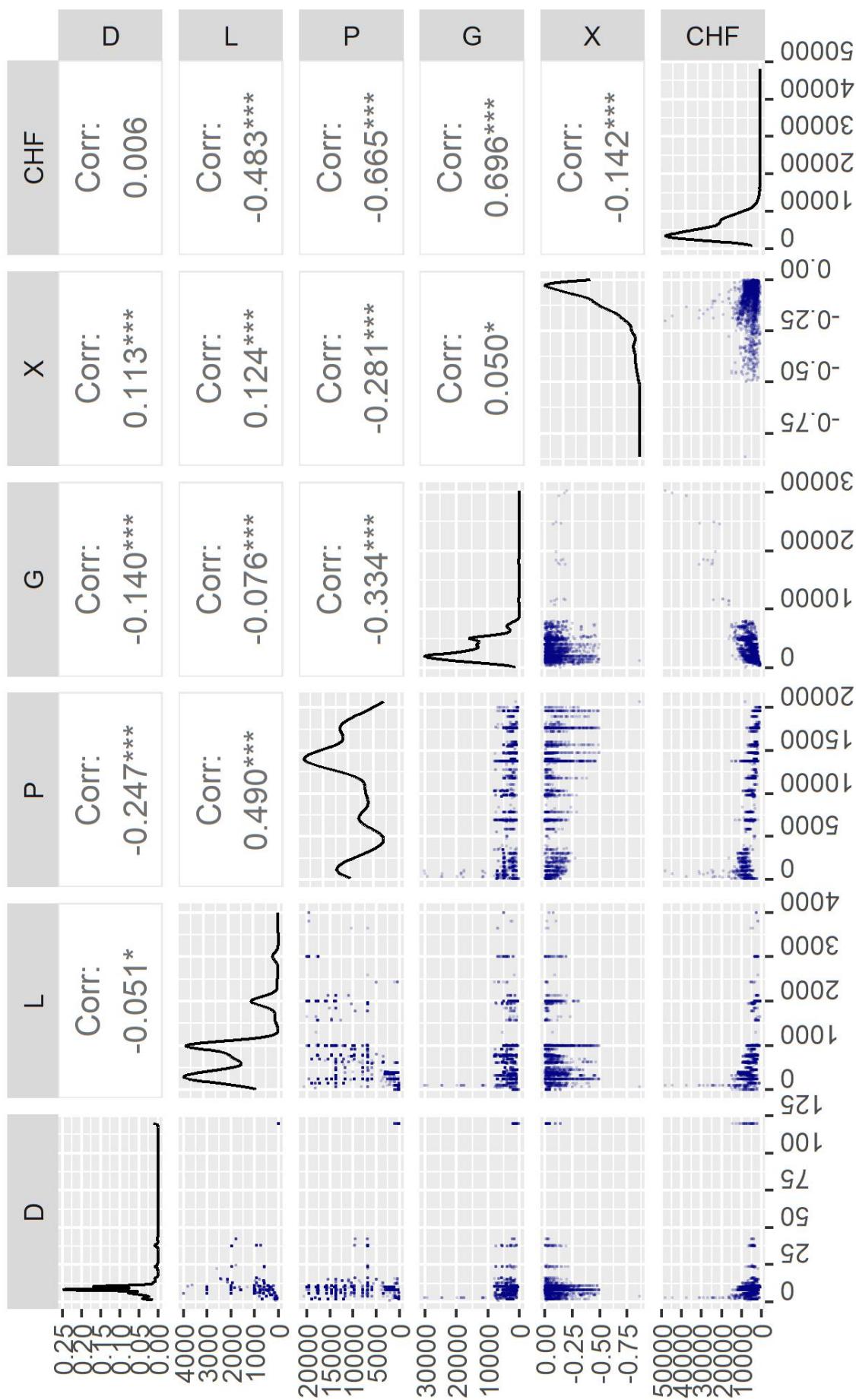


FIGURE 3.1 – Distribution of the modeling dataset, on the lower diagonal, a scatterplot of two variables is shown, the diagonal shows the density plot of one variable, and the upper diagonal presents Pearson's correlation between two variables.

results obtained by the fit of a QGAM model in different levels of quantiles will be shown and discussed.

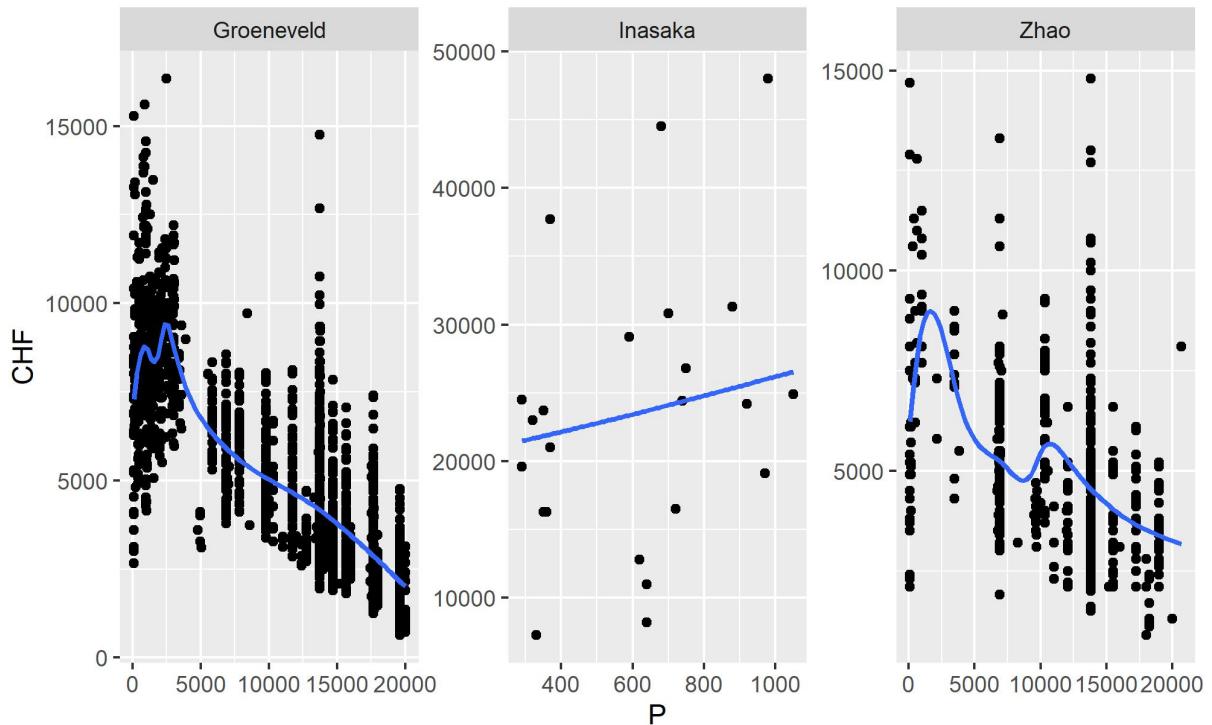


FIGURE 3.2 – Relationship between the pressure and the CHF for each data source, the blue line indicates a default GAM model available by the *ggplot2* library.

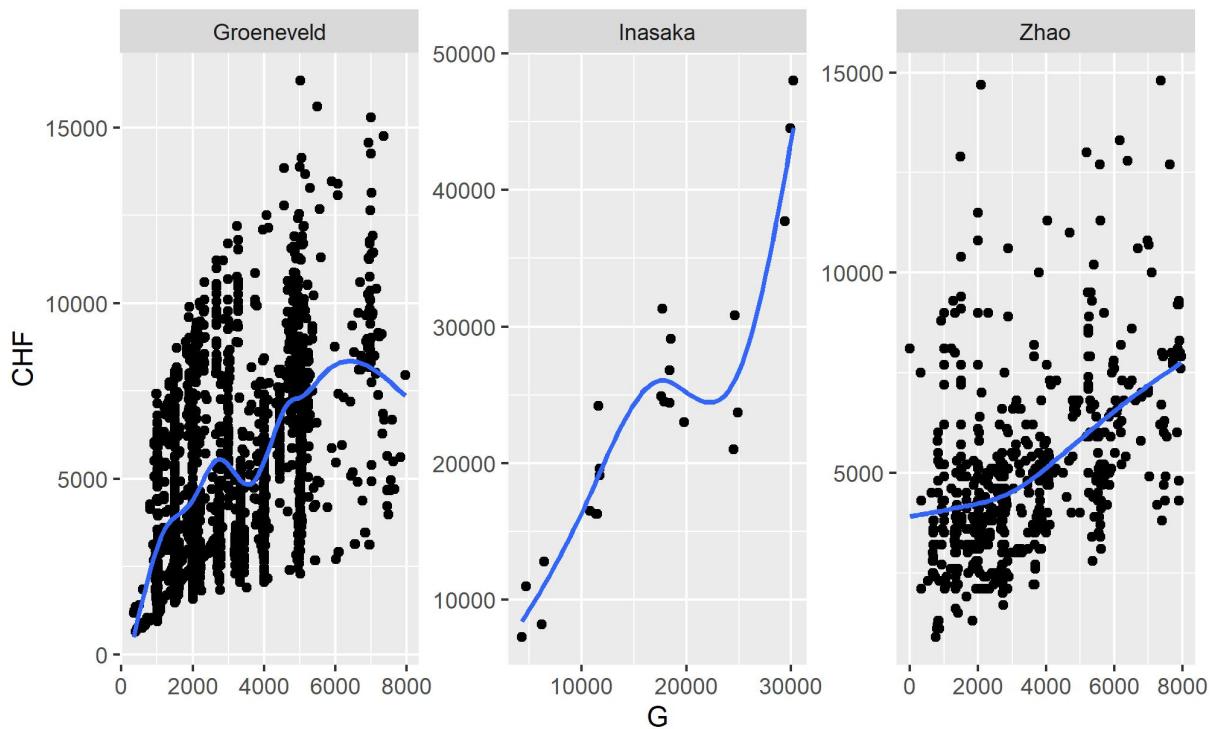


FIGURE 3.3 – Relationship between the mass flux and the CHF for each data source, the blue line indicates a default GAM model available by the *ggplot2* library.

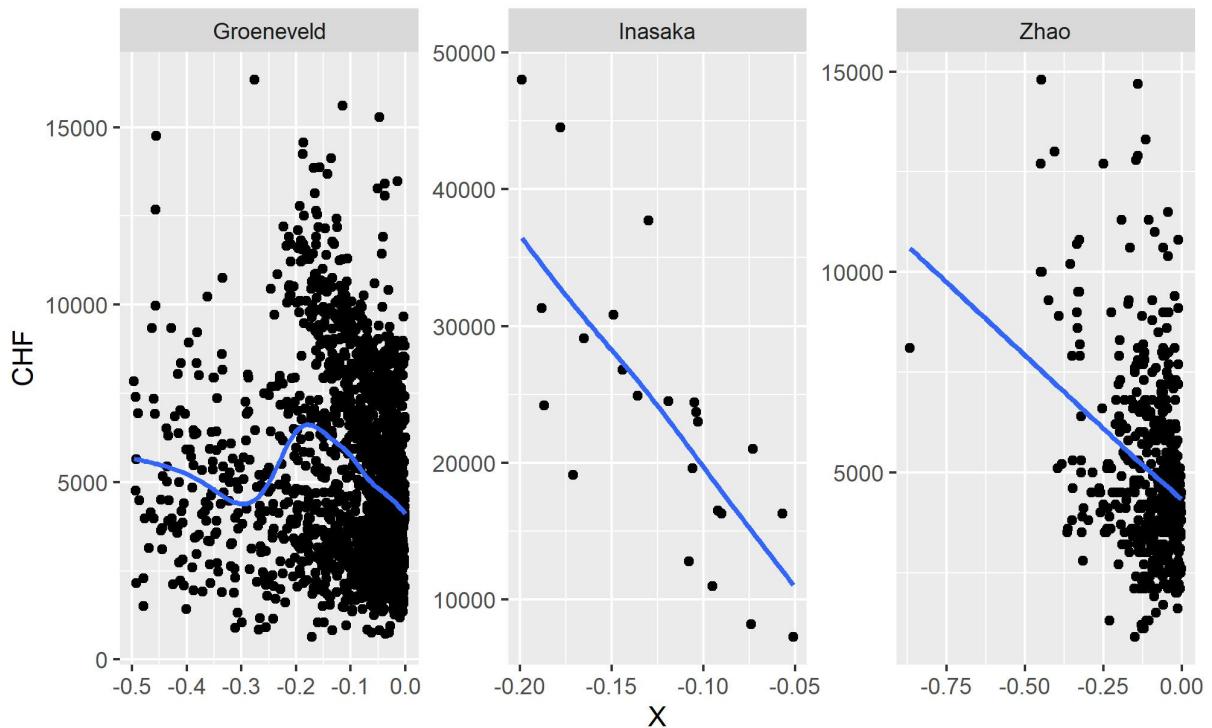


FIGURE 3.4 – Relationship between the thermodynamic quality and the CHF for each data source, the blue line indicates a default GAM model available by the *ggplot2* library.

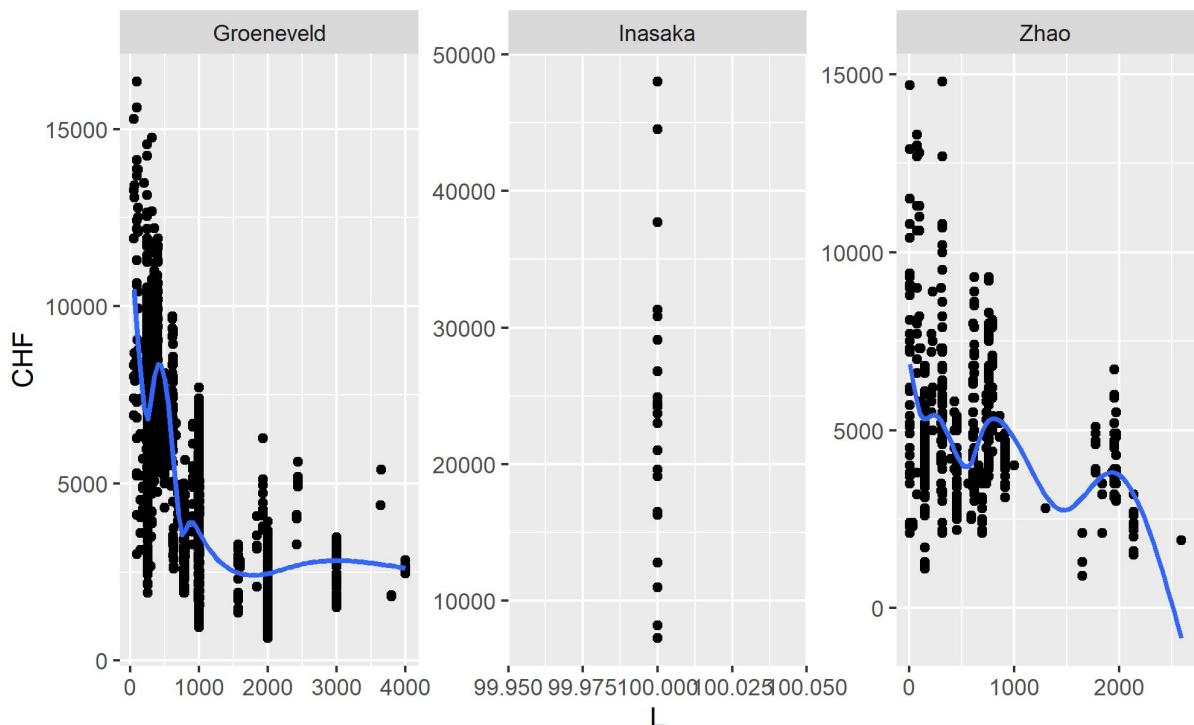


FIGURE 3.5 – Relationship between the heated length and the CHF for each data source, the blue line indicates a default GAM model available by the *ggplot2* library.

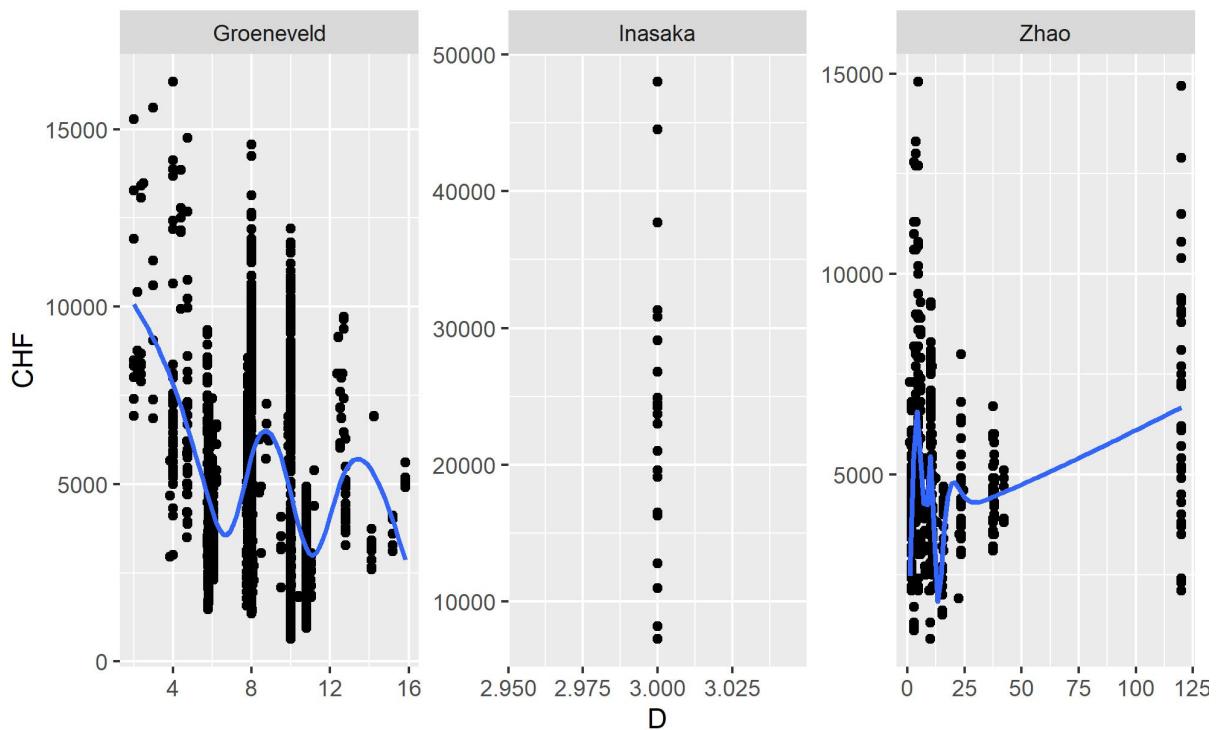


FIGURE 3.6 – Relationship between the diameter and the CHF for each data source, the blue line indicates a default GAM model available by the *ggplot2* library.

Figure 3.7 shows the distribution of the CHF for each data source. It is possible to observe that the measured CHF are all positive and right-skewed with a long tail, which indicates that a normal distribution of probabilities may not be the most suitable distribution for our use case. This project will test different distributions and link functions to find the one that best suits the modeling data.

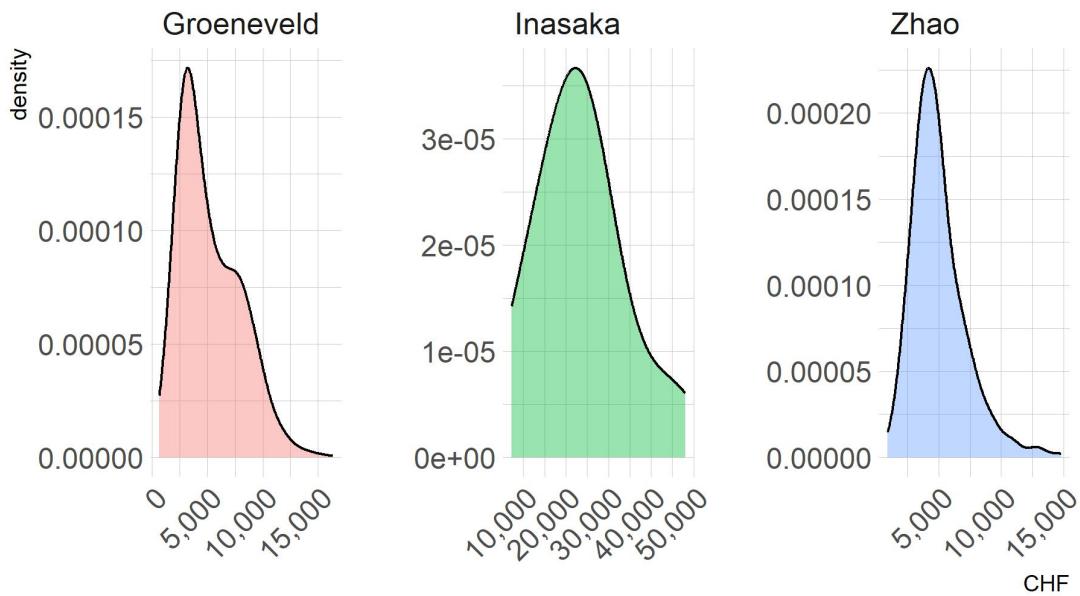


FIGURE 3.7 – Distribution of the CHF for each data source used in this work.

# 4 Methodology

## 4.1 Data Splitting

When dealing with statistical and machine learning models, using the complete dataset to fit the model is not recommended since it can give biased results as it uses all the data to adjust its parameters. In addition, when using the complete dataset to fit the model, it is not possible to measure how well it performs when new observations are given to it, making it harder to measure the predictive capacity and whether it is an overfitted or underfitted model. Usually, when one has a large dataset, the complete dataset follows a simple split into two random independent subsets: one for training the model (usually 70% of the data) and the other to test it (30% of the data).

In this work, the available dataset is not big enough to make a single split and evaluate the results to determine if one method got better predictive results than others due to the randomness of the single split. Therefore, we consider the following split approach. First, the data is split into two parts: a fitting dataset with 90% (2513 observations) and a validation dataset containing the other 10% of the data. The first split is applied to ensure that the validation data does not contain observations from Groeneveld's dataset, as the lookup table was developed with most of the literature's available data, which can make the final evaluation biased towards the lookup table. Then, the fitting data is split using the  $k$ -fold cross-validation (CV) methodology (JAMES *et al.*, 2023, pp. 206-209) to assess the fitted model predictive performance in different samples of the data.

The  $k$ -fold cross-validation methodology splits the entire fitting data into  $k$ -independent subsets. One of the  $k$  subsets is treated as the test dataset to measure the predictive quality. The other  $k - 1$  subsets are used to train the model. This process is repeated  $k$  times until all the  $k$  subsets are used as a test dataset. At the end of the process,  $k$  models were fitted, and a global metric of predictive quality can be measured. Figure 4.1 exemplifies the methodology when for  $k = 10$ .

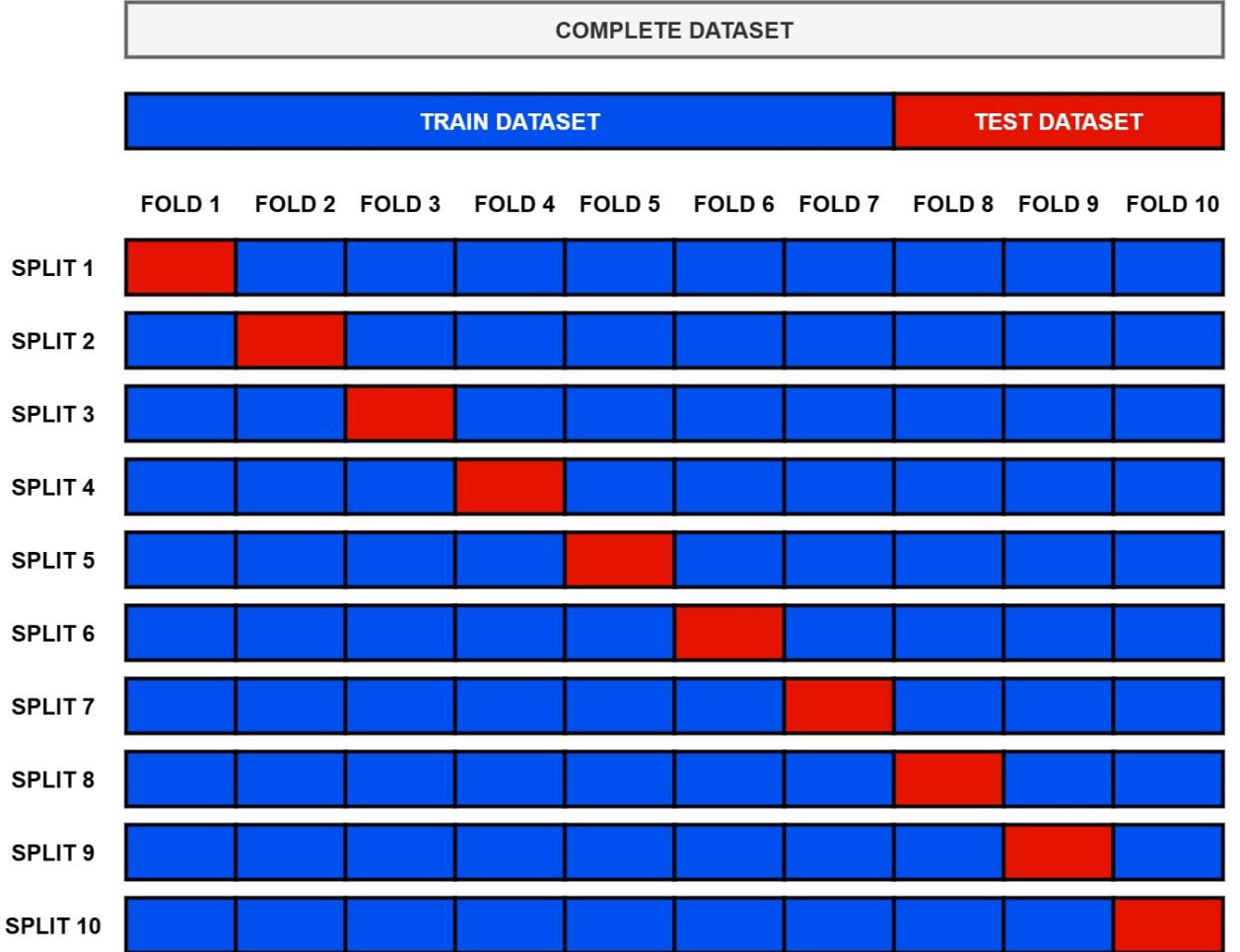


FIGURE 4.1 – Diagram of a simple splitting of the dataset and the  $k$ -fold cross-validation method for  $k = 10$ , the blue boxes indicate the data used to fit the model, the red ones indicate the data to be used as a test dataset to evaluate the predictive performance of the fitted model.

## 4.2 Evaluation Metrics

In this work, the prediction performance of each fitted model is evaluated for each split in the  $k$ -fold CV using metrics commonly applied in the machine learning literature and also in studies related to the prediction of CHF (HALL, 1999; GROSFILLEY, 2022). The evaluation metrics are presented as follows:

$$\text{RE}_i\% = \frac{CHF_{measured_i} - CHF_{pred_i}}{CHF_{measured_i}} \times 100 , \quad (4.1)$$

$$\text{ME}_k = \frac{1}{N} \sum \frac{CHF_{measured} - CHF_{pred}}{CHF_{measured}} \times 100 , \quad (4.2)$$

$$\text{MAE}_k = \frac{1}{N} \sum \frac{|CHF_{measured} - CHF_{pred}|}{CHF_{measured}} \times 100 , \quad (4.3)$$

$$\text{RMSE}_k = \sqrt{\frac{1}{N} \sum \left( \frac{\text{CHF}_{\text{measured}} - \text{CHF}_{\text{pred}}}{\text{CHF}_{\text{measured}}} \right)^2} \times 100 , \quad (4.4)$$

$$\text{ME}_{cv} = \frac{1}{K} \sum_{k=1}^{10} \text{ME}_k , \quad (4.5)$$

$$\text{MAE}_{cv} = \frac{1}{K} \sum_{k=1}^{10} \text{MAE}_k , \text{ and} \quad (4.6)$$

$$\text{RMSE}_{cv} = \frac{1}{K} \sum_{k=1}^{10} \text{RMSE}_k , \quad (4.7)$$

where  $i$  corresponds to the  $i$ th observation and  $N$  is the size of the test data set used to validate the model.

The relative error (RE) metric in Equation 4.1 is used to determine whether a model is overpredicting or underpredicting a CHF value compared to its observed value. Meanwhile, Equations 4.2, 4.3, and 4.4 define local metrics that assess model performance for each split in the  $k$ -fold cross-validation. The first metric (ME) quantifies the extent of overprediction or underprediction, the second (MAE) provides an absolute measure of how close the predictions are to the observed CHF values, and the third (RMSE) penalizes predictions that deviate significantly from observed values. In contrast, Equations 4.5, 4.6, and 4.7 represent global metrics, summarizing model performance across all cross-validation splits.

In addition, the validation dataset is crucial for evaluating how well each model performs when making predictions on new data. To evaluate the predictive performance on the validation dataset, the model fit from the  $k$ -split with the lowest  $\text{RMSE}_k$  in the test data is selected to generate predictions. Consequently, Equations 4.2, 4.3, and 4.4 are calculated without the index  $k$  in their formulas, and the subscript *val* is used for these metrics.

### 4.3 Model fitting

The models discussed in Chapter 2 are implemented using the R and Python programming languages, the package *mgcv* (WOOD, 2011) is used to fit the GAM model, the package *qgam* (FASIOLO *et al.*, 2021b) to fit the QGAM model, and the library *pyXSteam* to obtain parameters from steam tables, like the vapor density and surface tension.

Groeneveld's lookup table (GROENEVELD *et al.*, 2007) will have two fits varying the

parameter  $n$  shown in Equation 2.4, one considering  $n = 0.5$  and the other with  $n = 0.312$ . The Hall-Mudawar correlation (HALL; MUDAWAR, 2000b) has its parameters fixed by the author. However, both correlations (inlet and outlet conditions) are evaluated.

The additive models are assessed across different scenarios to evaluate which conditions they perform better under. For the parametric models, other distributions of the response variable, link functions, and formulas determining the relationship between  $Y$  and  $X$  are considered. For the QGAM model, different quantiles and formulas are assessed.

Table 4.1 summarizes all models considered in the model fitting process. The **R formula** column indicates how the parameters were given to the model to be fitted. The  $\sim$  symbol indicates that the equation models the response variable in the left-hand side of the equation. The  $s()$  symbol indicates the use of a smoothing function. Inside of  $s()$ , the “bs” means the type of basis function used. The basis functions tested in this work were adaptative smoothing (ad), thin plate regression splines (tp), and cubic regression splines (cr).

It is important to emphasize that the QGAM with quantile  $\tau = 0.95$  will not have its predictive results compared to the other models since these models aim to investigate their capability to fit relationships related to extreme values of CHF.

TABLE 4.1 – Additive models evaluated in this work, varying the distribution, link function, relationship structure, and quantile when possible

<b>Model</b>	<b>Distribution</b>	<b>Link function</b>	<b>R formula</b>	<b>Quantile</b>
LM	Gaussian	$\eta = \mu$	$\log(\text{CHF}) \sim P + G + X + L + D$	-
GLM 1	Gaussian	$\log(\mu) = \eta$	$\text{CHF} \sim P + G + X + L + D$	-
GLM 2	Gamma	$\log(\mu) = \eta$	$\text{CHF} \sim P + G + X + L + D$	-
GAM 1	Gaussian	$\log(\mu) = \eta$	$\text{CHF} \sim s(P,\text{bs}=\text{"ad"}) + s(X,\text{bs}=\text{"ad"}) + s(G,\text{bs}=\text{"ad"}) + s(D,\text{bs}=\text{"ad"}) + s(L,\text{bs}=\text{"ad"})$	-
GAM 2	Gamma	$\log(\mu) = \eta$	$\text{CHF} \sim s(P,\text{bs}=\text{"ad"}) + s(X,\text{bs}=\text{"ad"}) + s(G,\text{bs}=\text{"ad"}) + s(D,\text{bs}=\text{"ad"}) + s(L,\text{bs}=\text{"ad"})$	-
GAM 3	Gaussian	$\log(\mu) = \eta$	$\text{CHF} \sim s(P,\text{bs}=\text{"tp"}) + s(X,\text{bs}=\text{"tp"}) + s(G,\text{bs}=\text{"tp"}) + s(D,\text{bs}=\text{"tp"}) + s(L,\text{bs}=\text{"tp"})$	-
GAM 4	Gamma	$\log(\mu) = \eta$	$\text{CHF} \sim s(P,\text{bs}=\text{"tp"}) + s(X,\text{bs}=\text{"tp"}) + s(G,\text{bs}=\text{"tp"}) + s(D,\text{bs}=\text{"tp"}) + s(L,\text{bs}=\text{"tp"})$	-

GAM 5	Gaussian	$\log(\mu) = \eta$	$\text{CHF} \sim s(P, \text{bs}=\text{"cr"}) + s(X, \text{bs}=\text{"cr"}) + s(G, \text{bs}=\text{"cr"}) + s(D, \text{bs}=\text{"cr"}) + s(L, \text{bs}=\text{"cr"})$	-
GAM 6	Gamma	$\log(\mu) = \eta$	$\text{CHF} \sim s(P, \text{bs}=\text{"cr"}) + s(X, \text{bs}=\text{"cr"}) + s(G, \text{bs}=\text{"cr"}) + s(D, \text{bs}=\text{"cr"}) + s(L, \text{bs}=\text{"cr"})$	-
GAM 7	Gamma	$\log(\mu) = \eta$	$\text{CHF} \sim s(P, k=15, \text{bs}=\text{"cr"}) + s(X, \text{bs}=\text{"cr"}) + s(G, \text{bs}=\text{"cr"}) + s(D, \text{bs}=\text{"cr"}) + s(L, \text{bs}=\text{"cr"}) + \text{ti}(P, X, \text{bs}=\text{"cr"}) + \text{ti}(P, G, \text{bs}=\text{"cr"}) + \text{ti}(P, D, \text{bs}=\text{"cr"}) + \text{ti}(P, L, \text{bs}=\text{"cr"}) + \text{ti}(X, G, \text{bs}=\text{"cr"}) + \text{ti}(X, D, \text{bs}=\text{"cr"}) + \text{ti}(X, L, \text{bs}=\text{"cr"}) + \text{ti}(G, D, \text{bs}=\text{"cr"}) + \text{ti}(G, L, \text{bs}=\text{"cr"}) + \text{ti}(D, L, \text{bs}=\text{"cr"})$	-
QGAM 1	-	-	$\text{CHF} \sim s(P, \text{bs}=\text{"ad"}) + s(X, \text{bs}=\text{"ad"}) + s(G, \text{bs}=\text{"ad"}) + s(D, \text{bs}=\text{"ad"}) + s(L, \text{bs}=\text{"ad"})$	0.5
QGAM 2	-	-	$\text{CHF} \sim s(P, \text{bs}=\text{"tp"}) + s(X, \text{bs}=\text{"tp"}) + s(G, \text{bs}=\text{"tp"}) + s(D, \text{bs}=\text{"tp"}) + s(L, \text{bs}=\text{"tp"})$	0.5
QGAM 3	-	-	$\text{CHF} \sim s(P, \text{bs}=\text{"cr"}) + s(X, \text{bs}=\text{"cr"}) + s(G, \text{bs}=\text{"cr"}) + s(D, \text{bs}=\text{"cr"}) + s(L, \text{bs}=\text{"cr"})$	0.5
QGAM 4	-	-	$\text{CHF} \sim s(P, k=15, \text{bs}=\text{"cr"}) + s(X, \text{bs}=\text{"cr"}) + s(G, \text{bs}=\text{"cr"}) + s(D, \text{bs}=\text{"cr"}) + s(L, \text{bs}=\text{"cr"}) + \text{ti}(P, X, \text{bs}=\text{"cr"}) + \text{ti}(P, G, \text{bs}=\text{"cr"}) + \text{ti}(P, D, \text{bs}=\text{"cr"}) + \text{ti}(P, L, \text{bs}=\text{"cr"}) + \text{ti}(X, G, \text{bs}=\text{"cr"}) + \text{ti}(X, D, \text{bs}=\text{"cr"}) + \text{ti}(X, L, \text{bs}=\text{"cr"}) + \text{ti}(G, D, \text{bs}=\text{"cr"}) + \text{ti}(G, L, \text{bs}=\text{"cr"}) + \text{ti}(D, L, \text{bs}=\text{"cr"})$	0.95

# 5 Results

This chapter presents the results obtained by the fitted models and their predictive performance; in addition, the best GAM model will be discussed in more detail to show the interpretation tools available and how one can take advantage of this model not only from the predictive perspective but also from interpretation and risk control. Furthermore, the results obtained from the QGAM for  $\tau = 0.95$  will be discussed, and how it can be used to control extreme values for CHF.

## 5.1 Model Comparisons

Table 5.1 compares the fitted models using the metrics discussed in Section 4.2. The GAM 7 model obtained the lower  $RMSE_{val}$  of 24.87, followed by the QGAM 2 model with a  $RMSE_{val}$  of 34.87, the lookup table with  $n = 0.312$  obtained a  $RMSE_{val}$  of 43.7, and the Hall-Mudawar correlation for outlet conditions got the fourth best  $RMSE_{val}$  of 49.55. One supposition of the results above is that the models from Groeneveld and Hall-Mudawar are overfitted to work in the conditions presented in the data they used. In contrast, the additive models fitted were developed to get good predictive results in different situations than the data the model was fed.

When evaluating the  $RMSE_{cv}$  on the test dataset, the lookup table for  $n = 0.312$  got the lower value of 33.86, while the GAM 7 got a  $RMSE_{cv}$  of 57.24. However, it can result from a data shift from the train and test dataset of a specific fold (in this study, the 10th fold was impacting all additive models), in which the domain of the variables is different. As the lookup table does not need the application of the k-fold cross-validation, and most of the data in the fitting dataset comes from Groeneveld's source, it can indirectly benefit the lookup table model. When evaluating both the lookup table for  $n = 0.312$  and the GAM 7 model without the 10th fold, the obtained  $RMSE_{cv}$  is respectively 31.26 and 32.68, showing that the GAM 7 model can obtain good predictive results compared to the lookup table. Furthermore, the QGAM 2 model obtained a  $RMSE_{val}$  of 34.87, closer to the GAM 7 model, which can indicate that quantile models can also be an alternative to parametric additive models and physical correlations and be more robust when dealing

with outliers as the predicted value is the median.

When analyzing the  $ME_{cv}$  and  $ME_{val}$ , most models tend to underpredict the CHF except for the Hall-Mudawar correlation for outlet conditions. Similar interpretations of the RMSE results can be seen in the MAE metric.

TABLE 5.1 – Results of the models fitted in this work. The \*\* mark indicates the models with lower  $RMSE_{val}$  for the four categories of models: lookup table, Hall and Mudawar correlation, GAMs, and QGAMs. The † indicates the influence of the 10th fold in the  $RMSE_{cv}$  on the GAM 7 model, which without this fold, the  $RMSE_{cv}$  is 32.68.

MODEL	TRAIN DATASET			TEST DATASET			BEST SPLIT LOWER $RMSE_{cv}$	VALIDATION DATASET		
	$ME_{cv}$	$MAE_{cv}$	$RMSE_{cv}$	$ME_{cv}$	$MAE_{cv}$	$RMSE_{cv}$		$ME_{val}$	$MAE_{val}$	$RMSE_{val}$
LOOKUP TABLE N=0.5	-2.05	15.41	39.21	-2.09	15.43	37.02	6	-11.7	28.78	53.65
LOOKUP TABLE N=0.312	-0.62	12.25	34.98	-0.63	12.26	33.86	6	-6.51	19.1	43.7*
HALL-MUDAWAR CORRELATION INLET CONDITIONS	-32.16	45.11	67.15	-32.05	45.03	65.92	2	-49.89	62.62	79.13
HALL-MUDAWAR CORRELATION OUTLET CONDITIONS	9.03	20.4	45.15	9.08	20.44	44.31	3	2.4	24.55	49.55**
LM	-2.72	18.53	43.04	-1.86	20.39	44.86	6	-4.31	18.75	43.3
GLM 1	-7.44	20.26	45.01	-6.03	21.63	45.82	6	-7.96	20.43	45.2
GLM 2	-5.39	18.86	43.42	-4.39	20.76	45.21	3	-5.23	18.85	43.41
GAM 1	-2.98	12.19	34.92	-4.25	15.72	39.34	2	-2.7	12.88	35.88
GAM 2	-2.01	10.98	33.14	-4.89	17.34	41.28	1	-0.88	11.04	33.22
GAM 3	-2.68	11.32	33.64	-1.86	16.82	40.37	6	-2.95	11.47	33.87
GAM 4	-1.84	10.44	32.31	-1.62	17.83	41.58	6	-2.39	11.47	33.87
GAM 5	-2.33	11.31	33.63	-8.58	20.56	43.01	6	-3.47	12.53	35.4
GAM 6	-1.79	10.28	32.07	-3.13	18.06	41.95	5	-3.18	12.9	35.92
GAM 7	-0.99	7.36	27.12	-77.98	87.18	57.24 †	6	-1.59	9.92	31.49**
QGAM 1	-1.16	13.56	36.82	-2.65	19.08	43.18	6	-4.66	14.9	38.6
QGAM 2	-0.8	12.5	35.35	-3.2	19.57	43.81	6	-2.34	12.16	34.87**
QGAM 3	-0.95	12.81	35.79	-4.12	20.34	44.28	6	-3.17	12.62	35.53

Figure 5.1 compares the measured CHF and the predicted CHF in the validation dataset. A good model fit is expected to result in a plot where dots are close to the red dashed line. The lookup table and Hall-Mudawar models have four highly underpredicted observations, reflected in the high metric values presented in Table 5.1. The GAM 7 and QGAM 2 models have their predictions closer to the dashed red line, indicating their better predictive power in the validation data than other methods, as shown in Table 5.1.

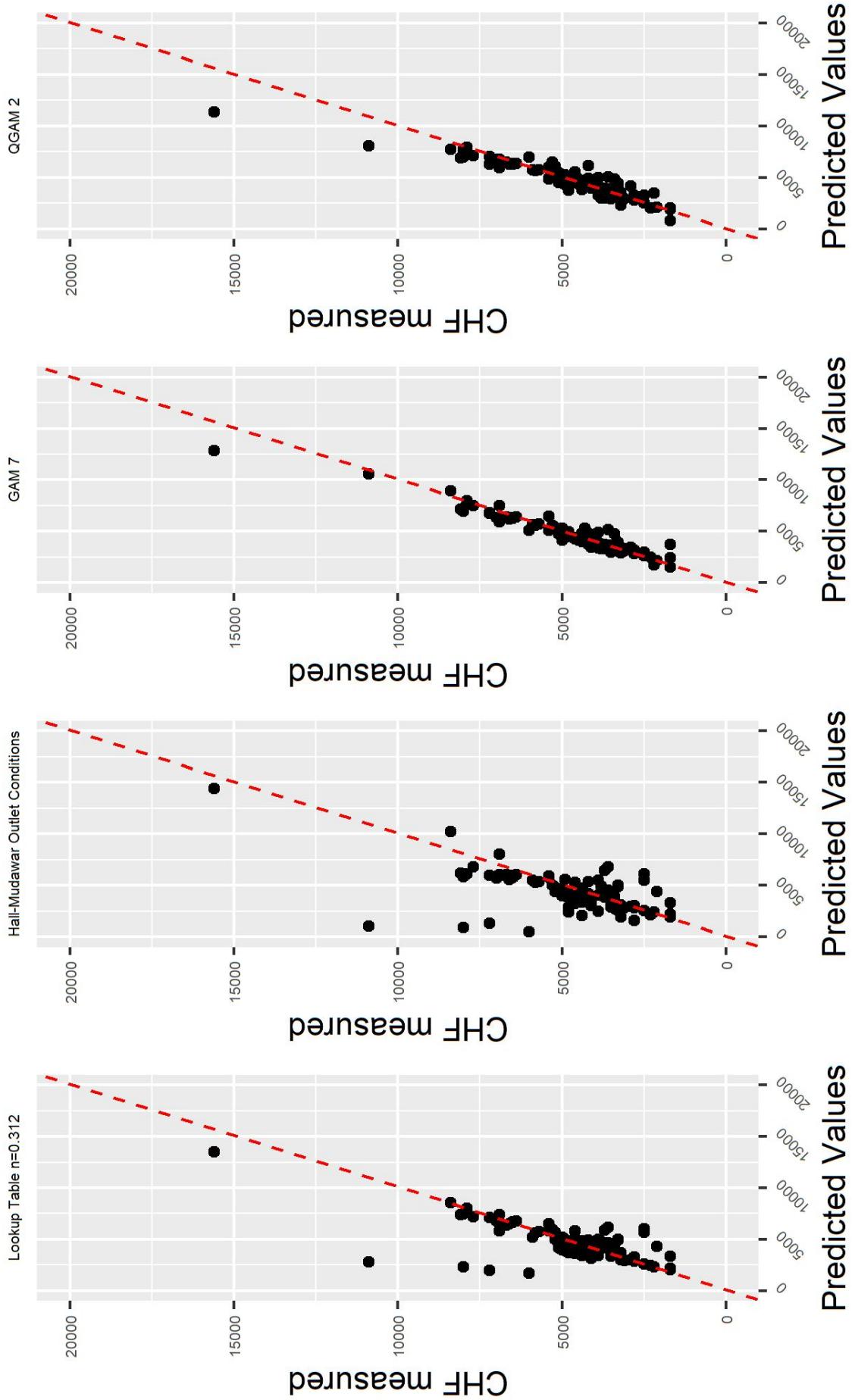


FIGURE 5.1 – Measured versus predicted CHF values by the models with the \*\* mark on Table 5.1

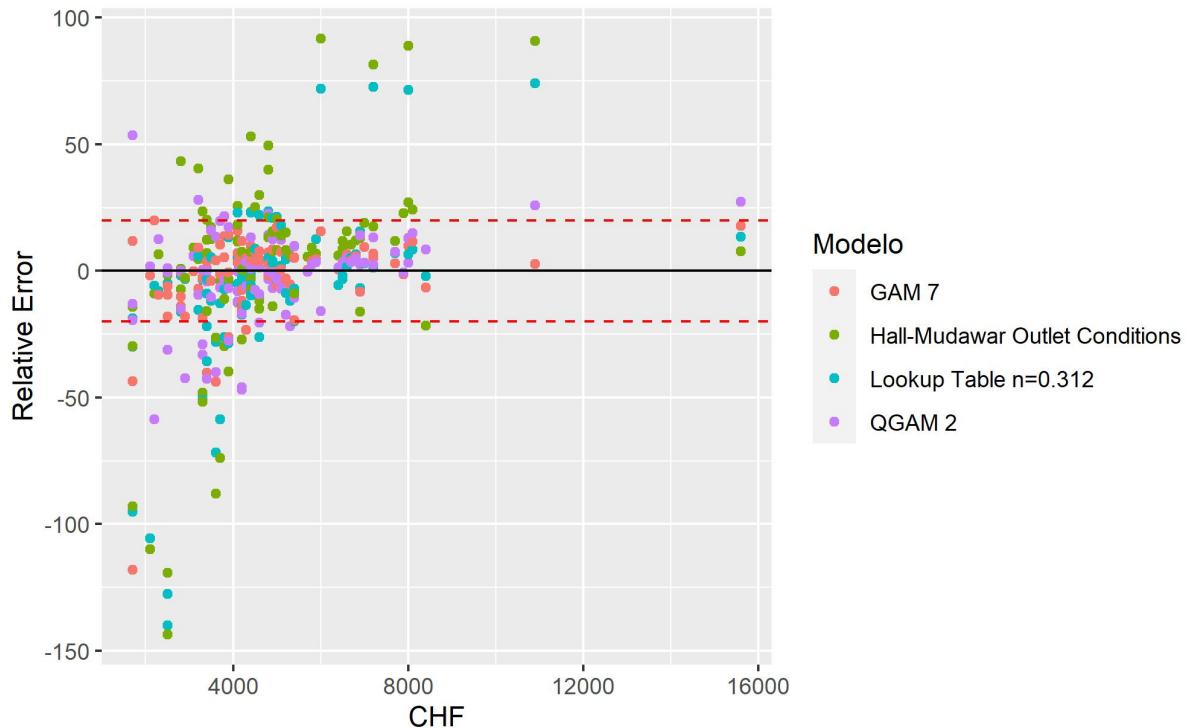


FIGURE 5.2 – Model's Relative Error for the measured CHF on the validation dataset.

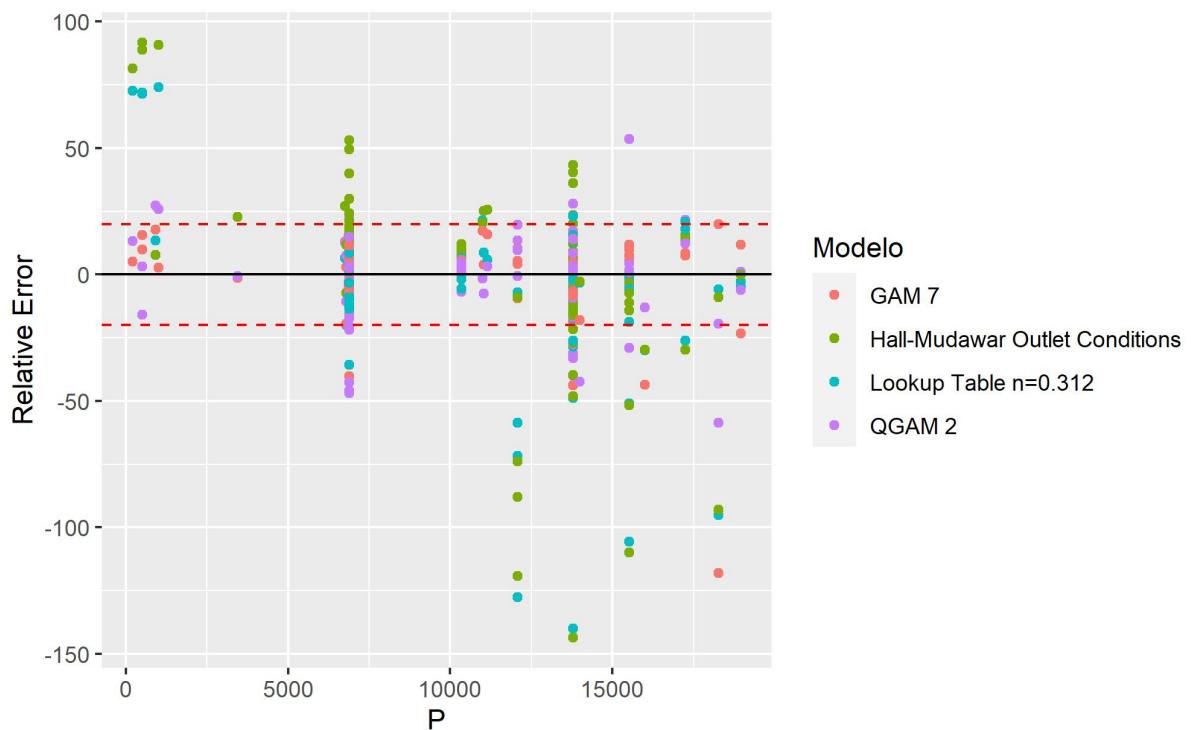


FIGURE 5.3 – Model's Relative Error for the measured pressure on the validation dataset.

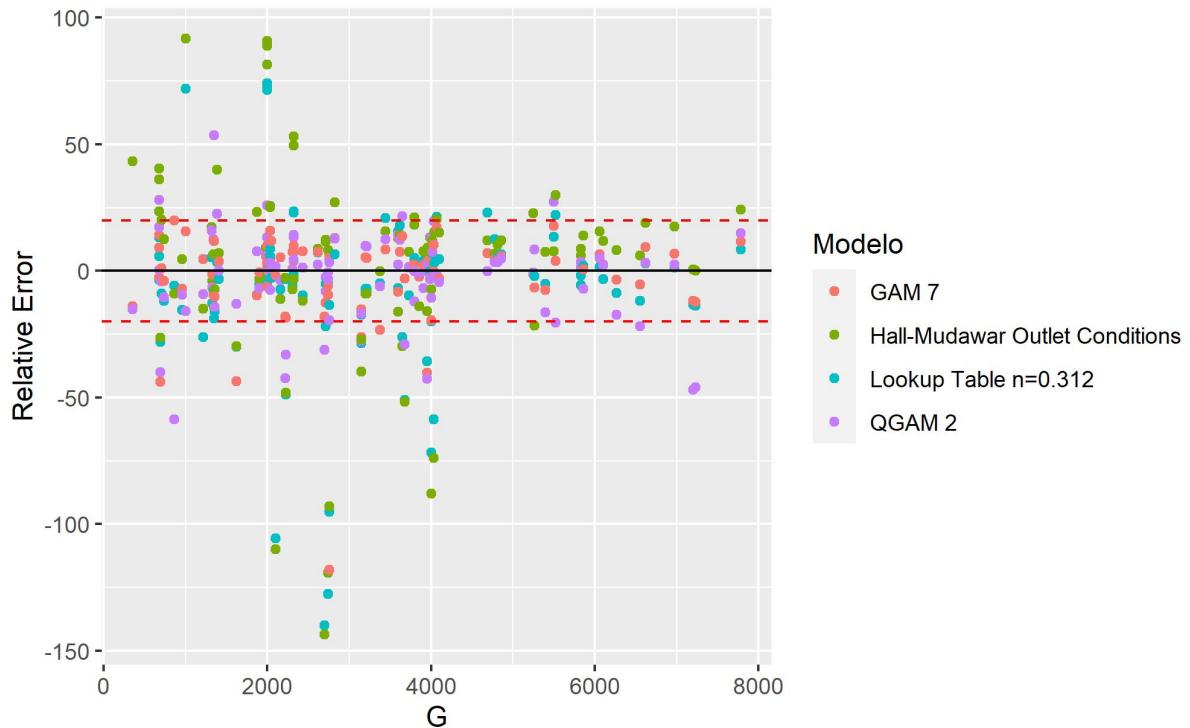


FIGURE 5.4 – Model's Relative Error for the measured mass flux on the validation dataset.

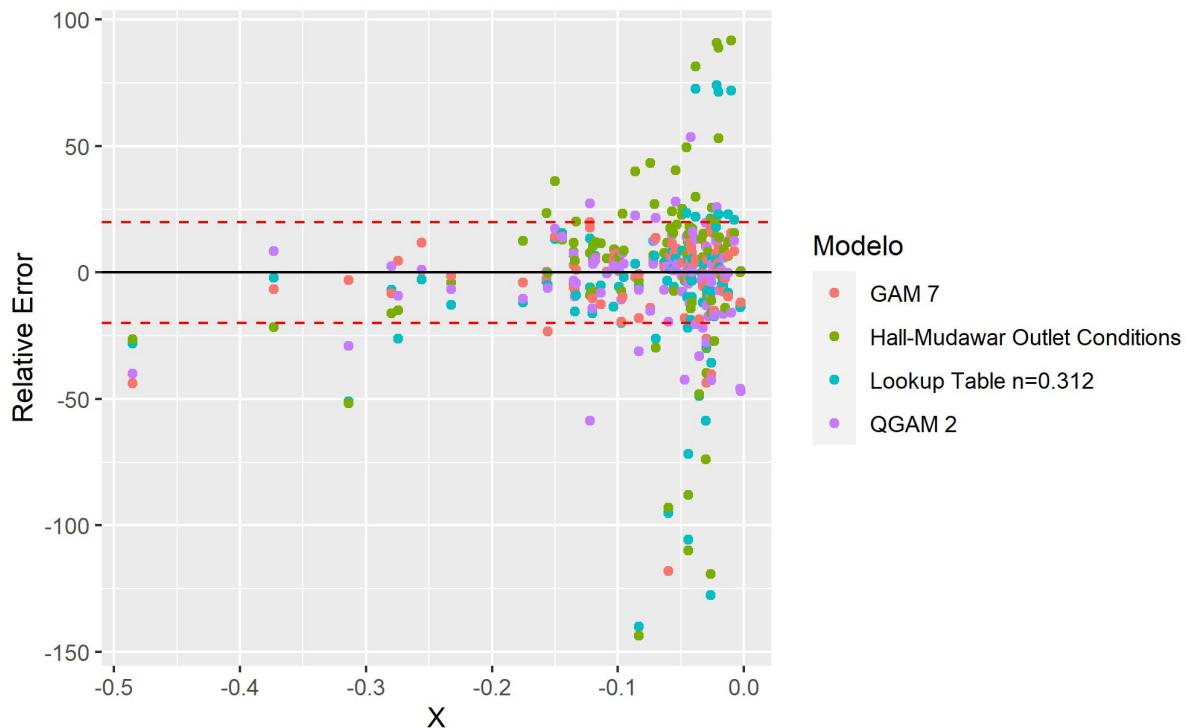


FIGURE 5.5 – Model's Relative Error for the measured thermodynamic quality on the validation dataset.

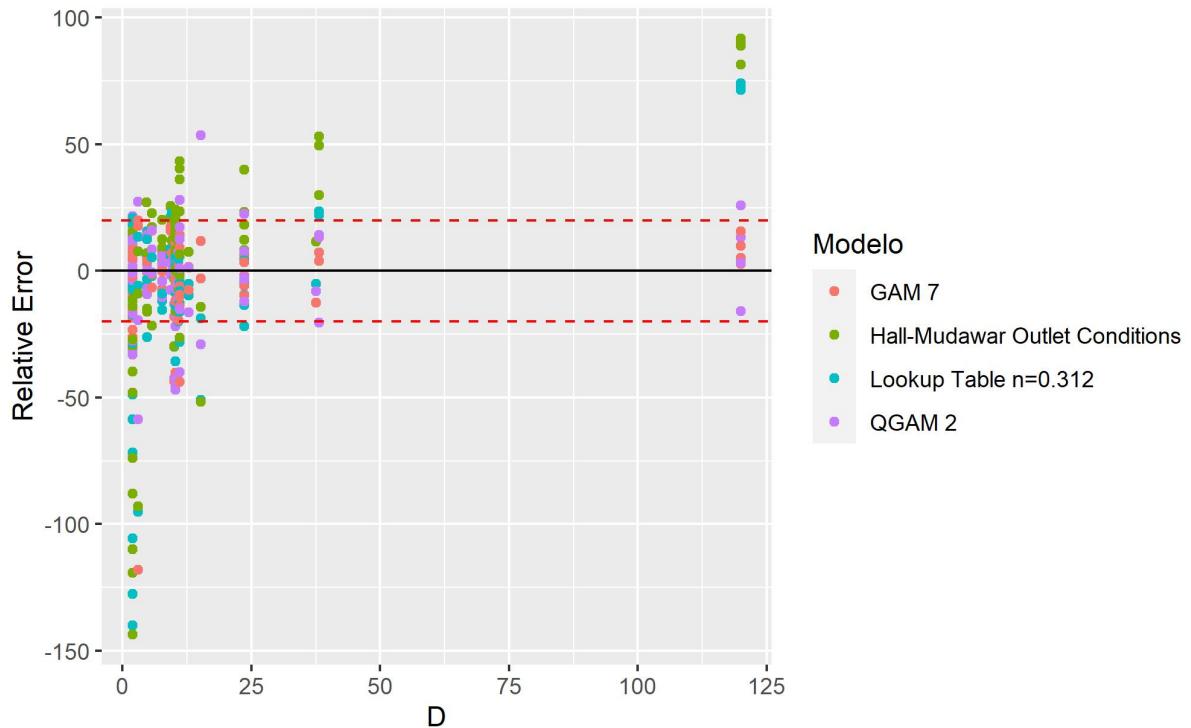


FIGURE 5.6 – Model's Relative Error for the measured heated diameter on the validation dataset.

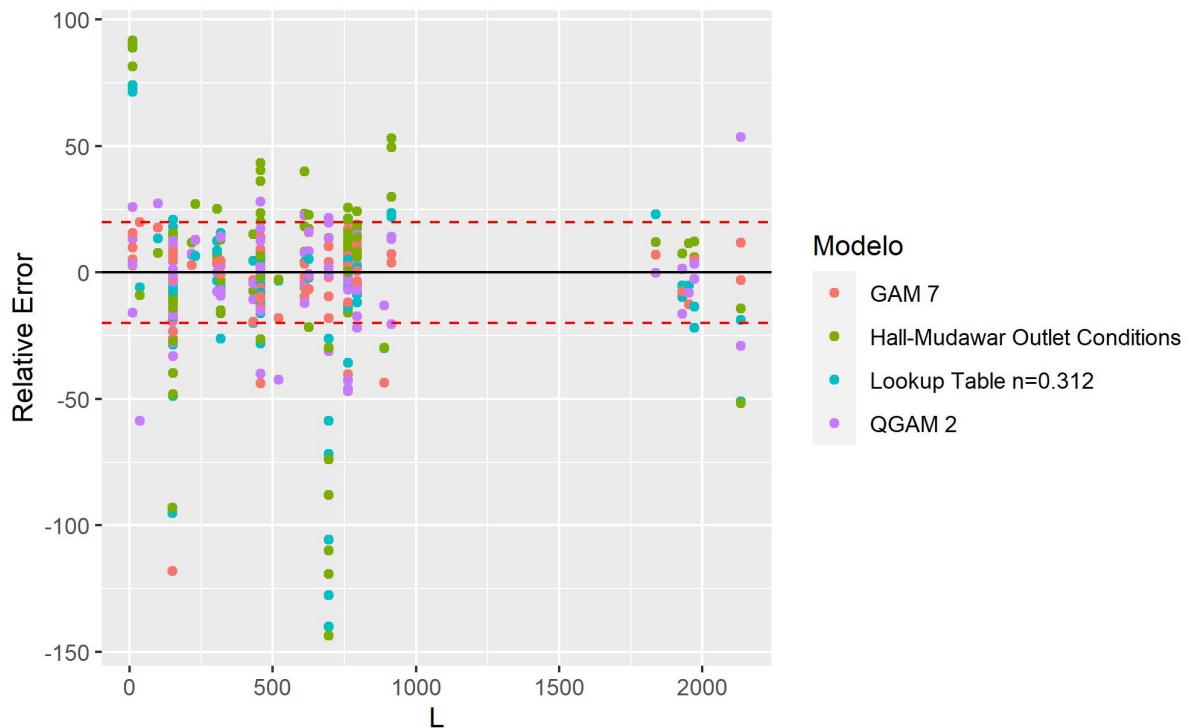


FIGURE 5.7 – Model's Relative Error for the measured heated length on the validation dataset.

Figures 5.2, 5.3, 5.4, 5.5, 5.6, and 5.7 show the  $RE_i\%$  (Equation 4.1) for the CHF and each input variable in the validation dataset; these plots try to identify in which regions of data each model fails to predict the CHF correctly. The red dashed line indicates the  $RE_i$  of  $-20\%$  and  $20\%$ . It is possible to observe that in the observations where the pressure  $P$  is lower, the Lookup Table and the Hall-Mudawar model overestimate the CHF. On the other hand, with pressure values greater than  $10000 \text{ kPa}$ , these models tend to underestimate the measured critical heat flux. When analyzing the thermodynamic quality  $X$  with values greater than  $-0.1$ , the Lookup Table and the Hall-Mudawar models have their relative errors increased. In addition, in this range of values, the QGAM 2 has six observations in which the CHF is underpredicted. On the contrary, lower values of the heated diameter  $D$  strongly influence the Hall-Mudawar and the Lookup Table predictions.

When analyzing Figures 5.2, 5.3, 5.4, 5.5, 5.6, and 5.7, the GAM 7 and the QGAM 2 models tend to be closer to the zero relative error than the Lookup table and the Hall-Mudawar correlations. In addition, there is no clear evidence of the influence of the inputs in the measured relative error in the GAM 7 and QGAM 2 models. This corroborates the evaluations made in Table 5.1 and Figure 5.1, in which the additive models can be considered a good alternative to the established methods in the literature discussed in this dissertation.

In the next section, the tools available to understand the expected behavior of the best GAM 7 model fitted in the k-fold cross-validation and how one can benefit from these results.

## 5.2 Best GAM Model Interpretation

One advantage of both GAM and QGAM models can be observed in Figures 5.8, 5.9, 5.10, 5.11, 5.12 and 5.13, in which a plot of the fitted smoothing functions can be obtained. In Figures 5.8, 5.9, 5.10, 5.11, and 5.12, the curve of each one-dimensional smoothing function fitted and the respective confidence intervals for the best GAM 7 model is shown, indicating that for the best GAM model fitted, high values of diameter, heated length, and mass flux, and low values of thermodynamic quality, predictions and inferences made with this model is not recommended, since the confidence intervals in these areas are bigger.

Figure 5.13 illustrates the two-dimensional tensor product functions fitted by the model, in which the grey area in the plot indicates regions of low confidence to make predictions as the data available in the fitting process was not big enough to make them reliable to make inference and prediction. Another characteristic of this type of contour plot is presented by the colors, in which the yellow represents higher values of the output

smoothing functions, and the blue color is associated with lower output values.

In Figures 5.8, 5.9, 5.10, 5.11, 5.12, and 5.13, the output from the fitted smooth function can be interpreted and how it affects the predicted CHF. For example, pressure values lower than 5,000  $kPa$  lead to higher values in the univariate smooth function and, consequently, a higher predicted CHF.

Furthermore, the importance of the GAMS models can be observed in the univariate curve of the pressure in Figure 5.8, in which the influence of this variable could not be explained by the use of a linear term as in the LM model fitted, as its behavior changes by the range of observed values.

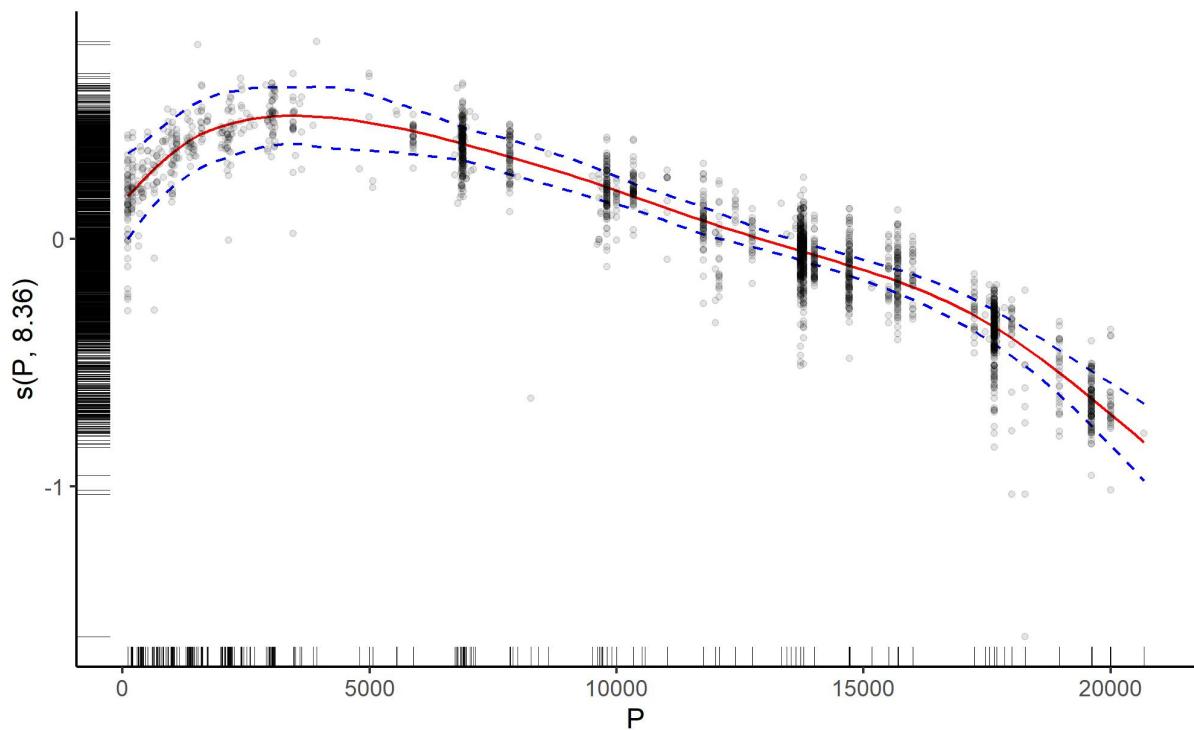


FIGURE 5.8 – One-dimensional smoothing function ( $s("P")$ ) fitted by the best GAM model in predictive performance for the pressure.

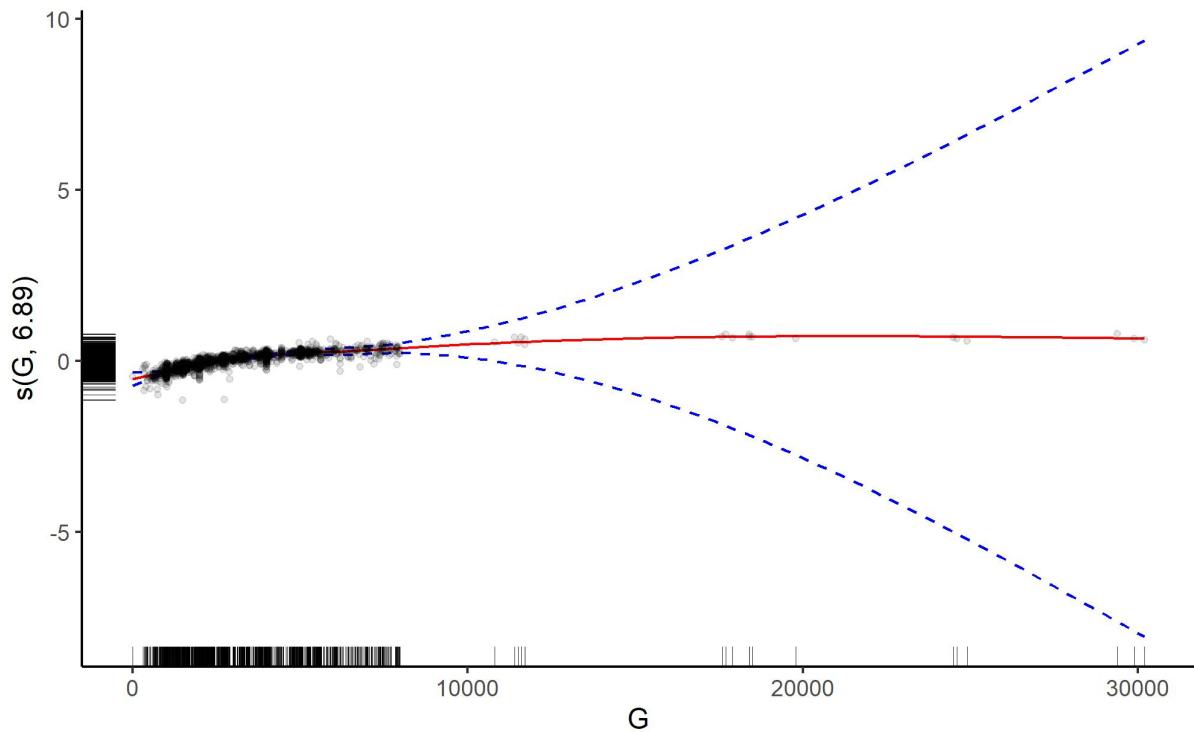


FIGURE 5.9 – One-dimensional smoothing function ( $s("G")$ ) fitted by the best GAM model in predictive performance for the mass flux.

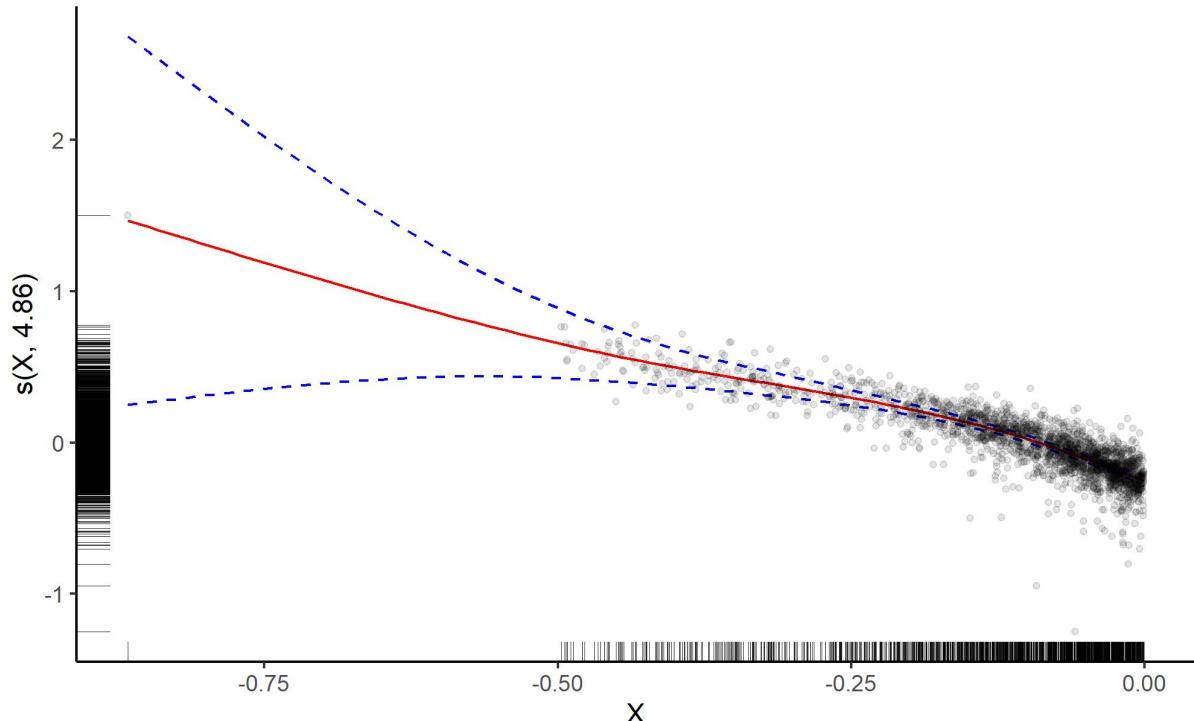


FIGURE 5.10 – One-dimensional smoothing function ( $s("X")$ ) fitted by the best GAM model in predictive performance for the thermodynamic quality.

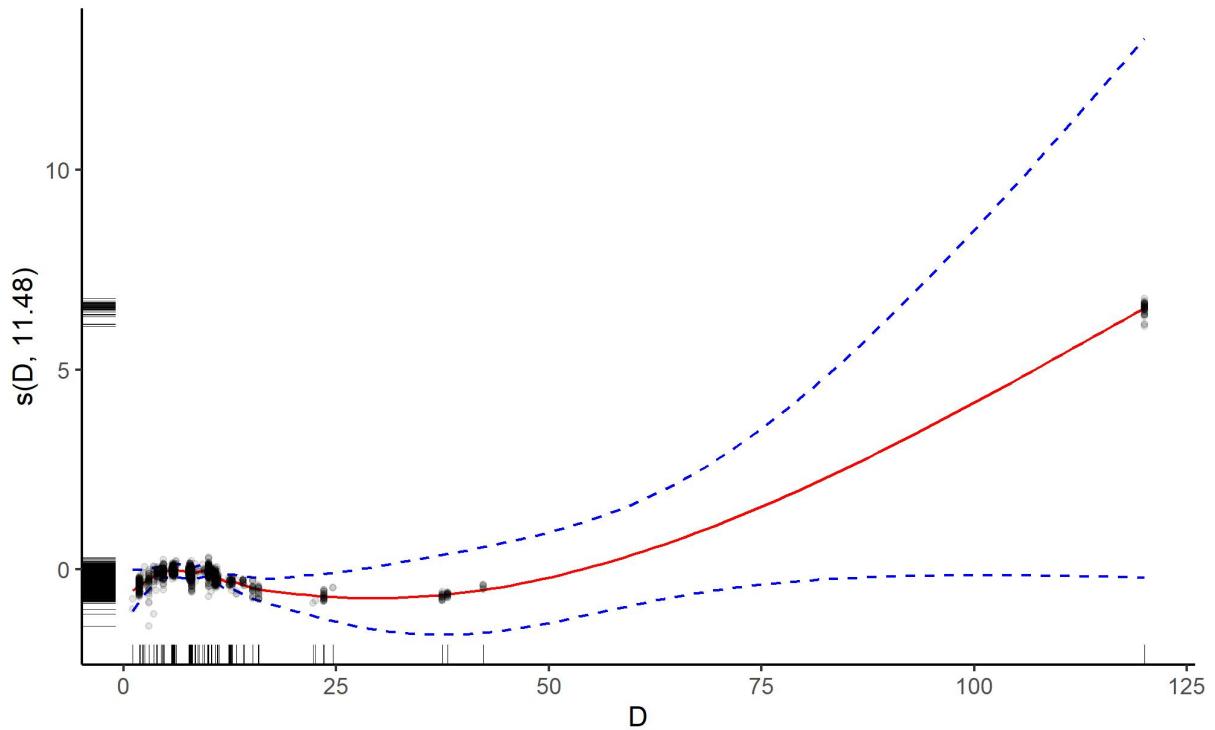


FIGURE 5.11 – One-dimensional smoothing function ( $s("D")$ ) fitted by the best GAM model in predictive performance for the heated diameter.

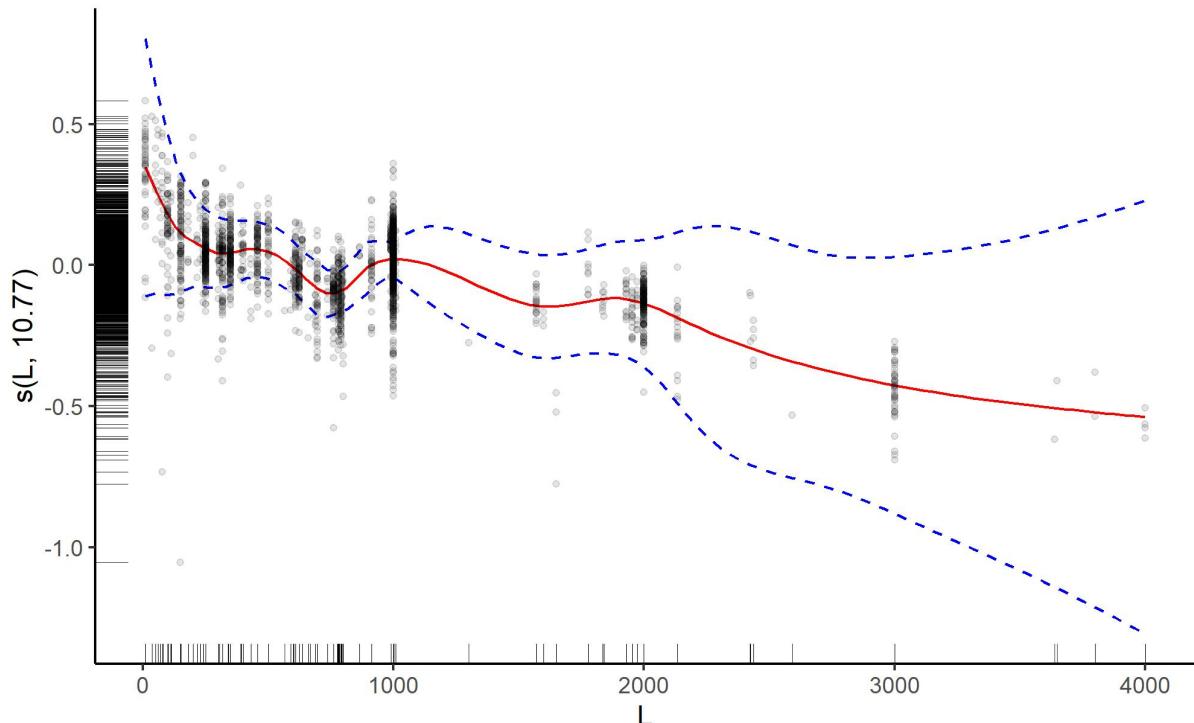
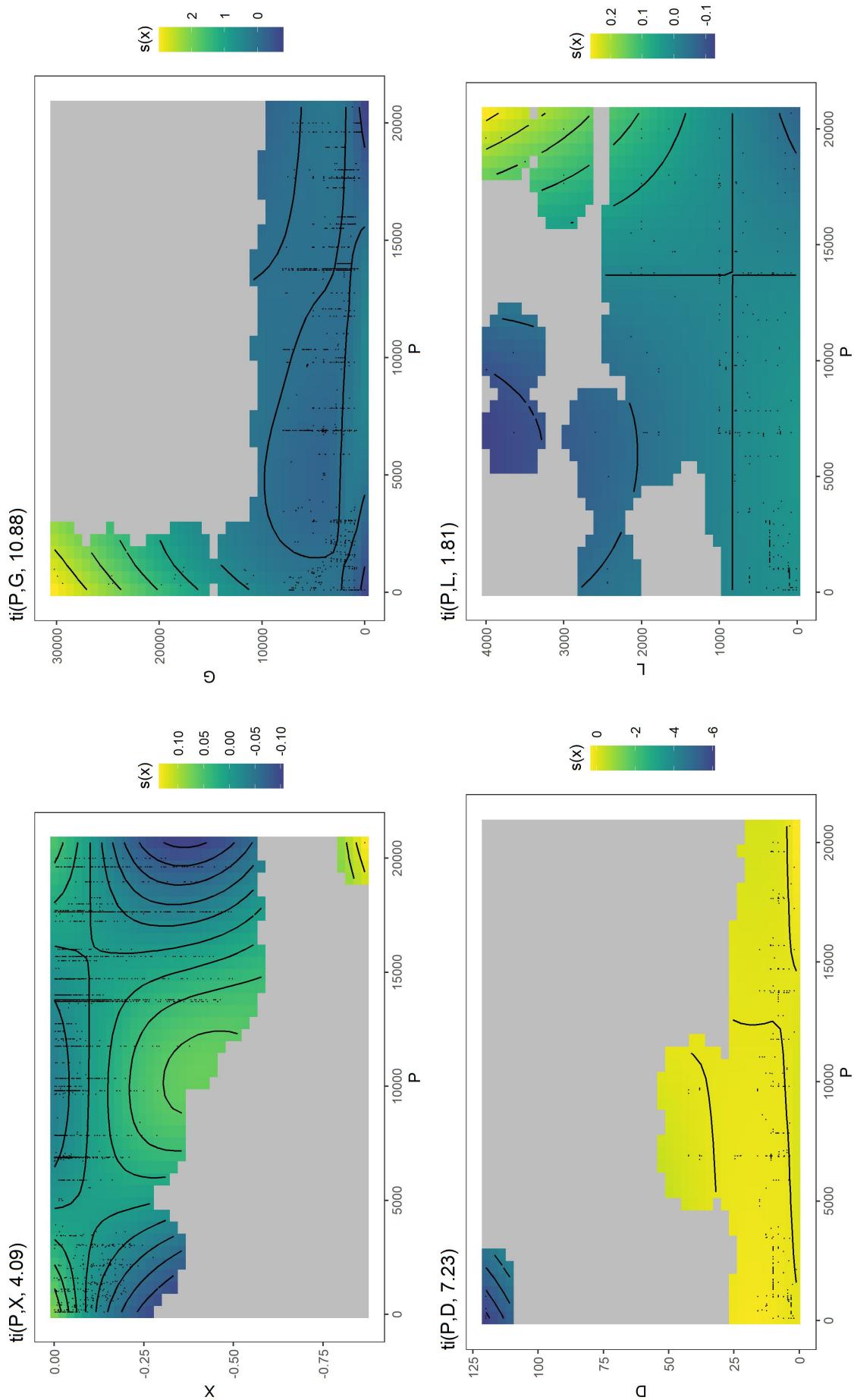


FIGURE 5.12 – One-dimensional smoothing function ( $s("L")$ ) fitted by the best GAM model in predictive performance for the heated length.



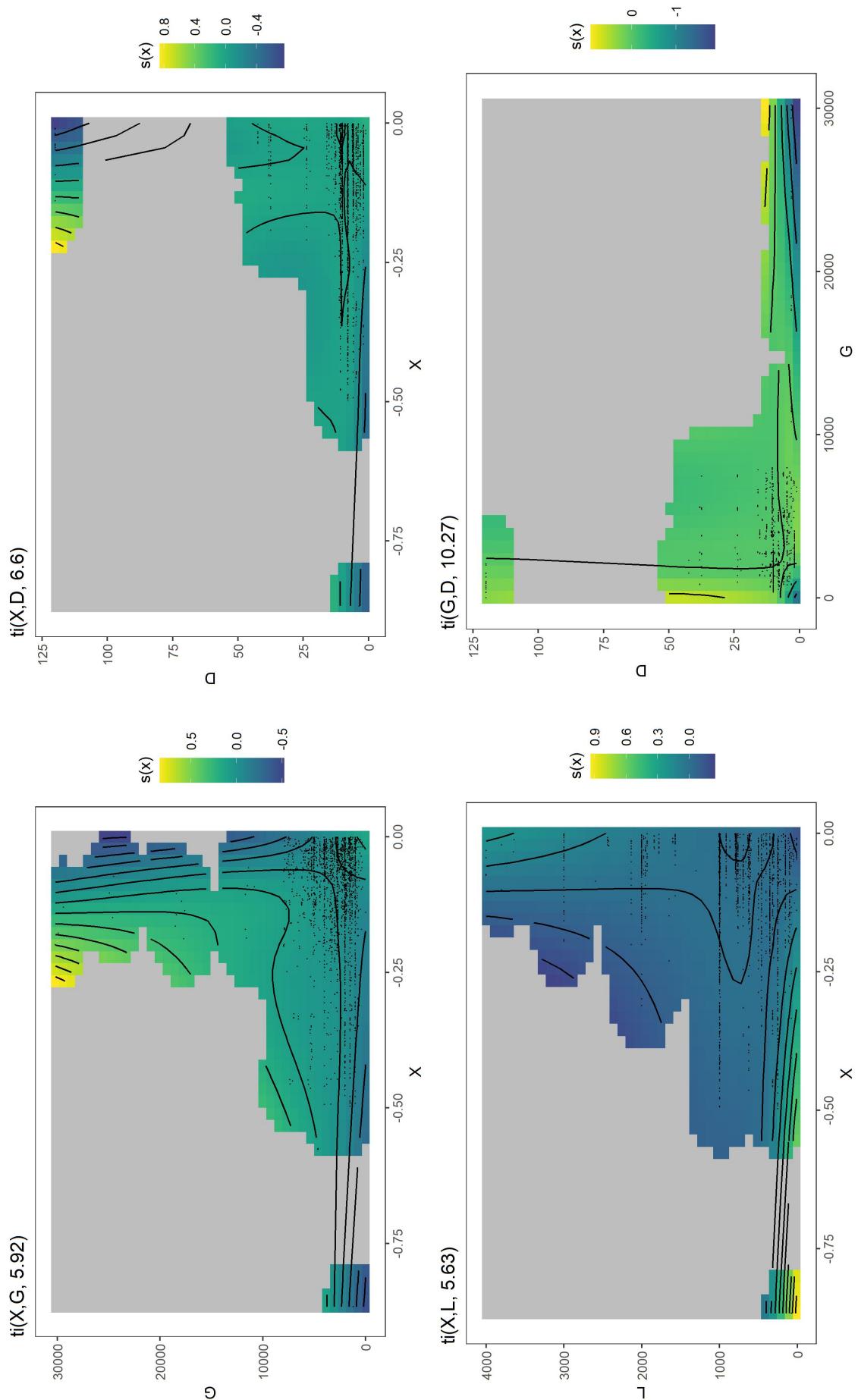
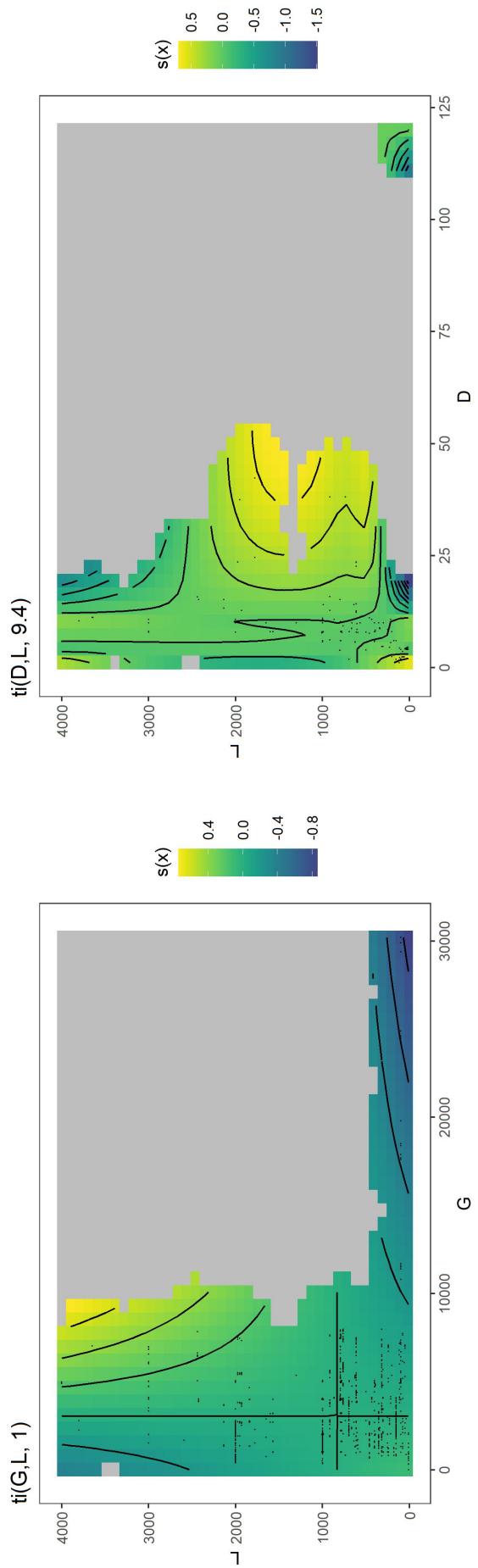


FIGURE 5.13 – Two-dimensional tensor functions ( $t_i()$ ) fitted by the best GAM model in predictive performance.



In addition, the GAM and QGAM models can return the respective prediction interval for each prediction. This enables users to evaluate the range of possible values the measured CHF may occur by a 95% interval, as illustrated in Figure 5.14.

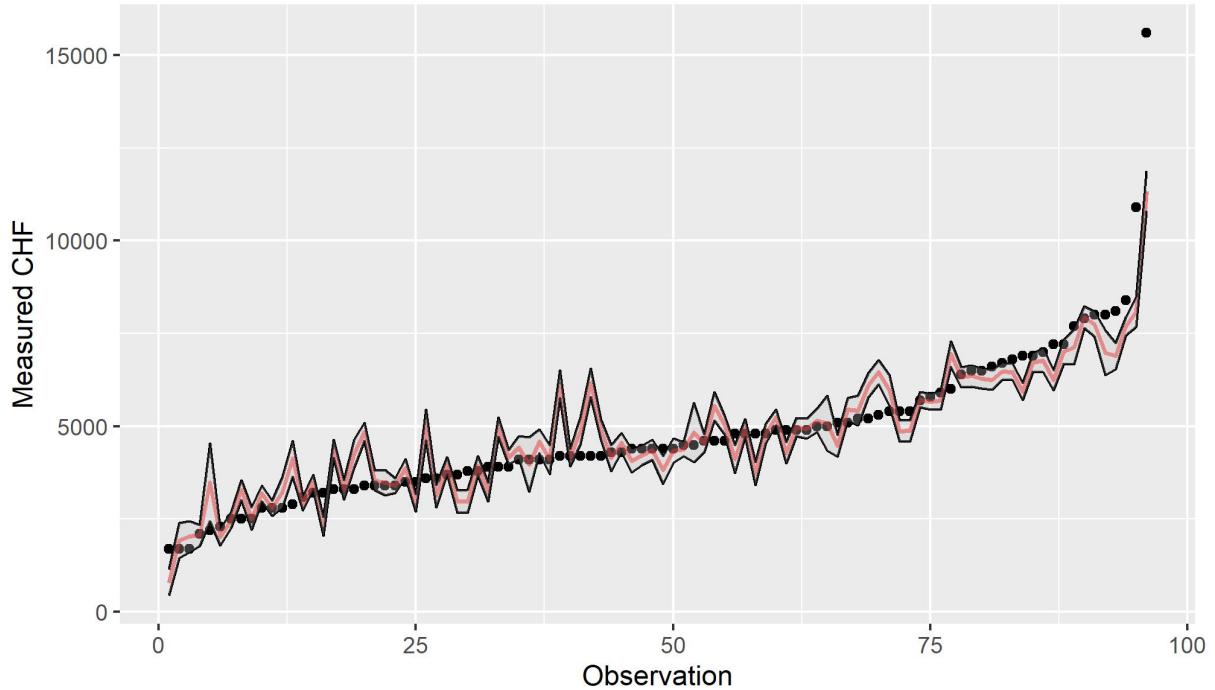


FIGURE 5.14 – Approximate Predictions Interval for the best GAM 7 model fitted in the validation dataset. The dots represent the measured CHF, and the red line is the predicted CHF. The black lines and the grey area represent the approximate 95% prediction interval.

As the GAM model is a method in which some assumptions are made when one wants not only to make predictions but also make inferences about the fitted model as discussed in Section 2.3, an analysis of the residuals is also available, as shown in Figure 5.15. The top-left plot shows the q-q plot of the deviance residuals, which measures if the chosen probability distribution and link function suit the data, in which a perfect model will have the residuals around the red line. The q-q plot in the GAM model fitted indicates that for the extreme values of CHF, both low and high values, the residuals deviate from the line, which can suggest that the training dataset has extreme values that differ from most of the data or the distribution used is not the best to explain the data, which shows that there is still room for improvement in the work, as new data and different probability distributions can be tested. The top-right plot shows the histogram of the residuals, which is desired to be zero-centered; it is possible to observe that it has a long tail towards negative residuals, indicating a possible outlier in the training data. The bottom-left plot shows the linear predictor against the residuals, which is useful for the evaluation of the residuals and the search for heteroscedasticity; in the GAM 7 model, this behavior can not be observed. The bottom-right plot is similar to Figure 5.1. However, it compares the fitted values in the model with the measured CHF values in the training data set; this plot for the GAM

model suggests that the fitted model does not have a trend or area in which the model fails to predict.

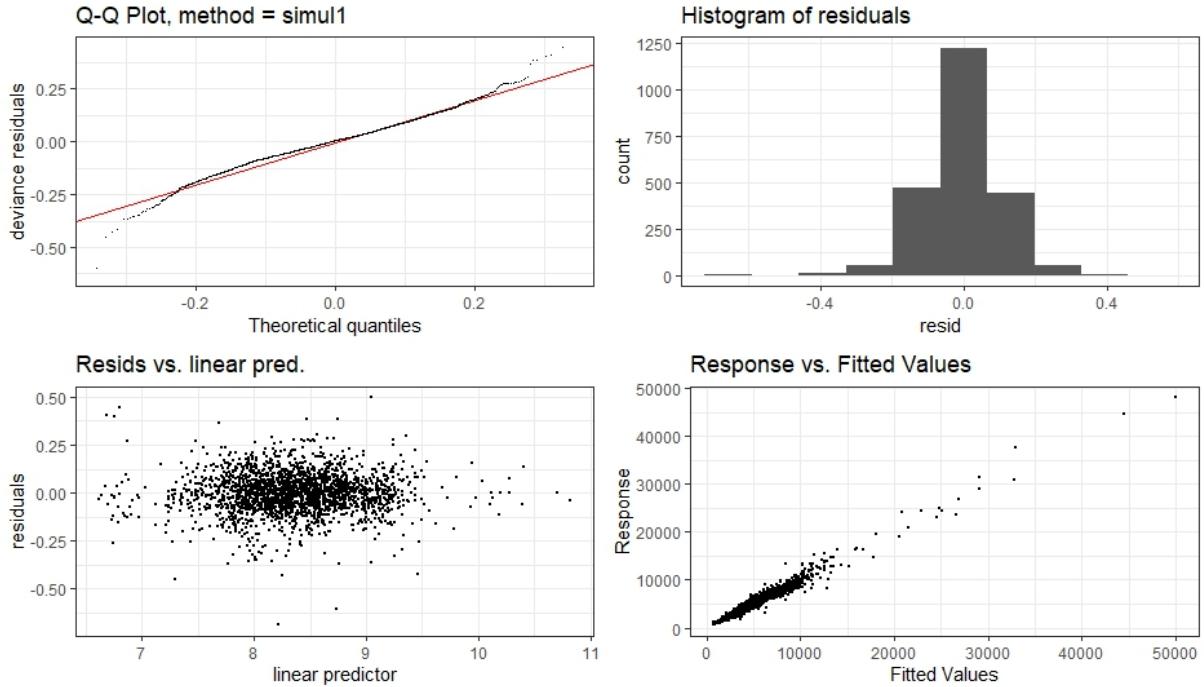


FIGURE 5.15 – Residual analysis of the best GAM 7 model fitted.

This section presented part of the available tools and the results from the best GAM 7 model fitted. Besides being the model with the best prediction quality in this work, it was observed that there is still room for improvement for the model, either by the test of new probability distributions which can take into account the assumptions necessary to be considered a GAM model or by the increase of the number of datasets by the collection of new data from other sources not used in this work.

The next section will similarly discuss how one can use the QGAM model to study and predict possible extreme values of the CHF and how to use it as a real-time alert system.

### 5.3 QGAM for Extreme Values

The QGAM 4 model shown in Table 4.1 was fitted by a different process than the previous models discussed, as our main objective in this model is to predict the CHF value in which the probability of observing a higher value is 5%, indicating a low probability event, that depending of the use case, for example on nuclear reactors, can be interpreted as a high-risk value. The QGAM 4 model fitted used the whole modeling dataset shown in Section 3.4 as the main objective of this model is to cover a wide range of CHF to

estimate the quantile  $\tau = 0.95$ .

Figures 5.16, 5.17, 5.18, 5.19, 5.20, and 5.21 show the fitted smoothing and tensor functions for the QGAM 4 model. For the univariate smoothing functions, the red line represents the estimated quantile  $\tau = 0.95$ , while the dashed blue lines represent the estimated confidence intervals of each smoothing function; one may notice that the behavior of these functions is similar to the ones from the GAM 7 model, in which high values of mass flux, heated length, and diameter are not recommended to make predictions or inference, in addition, the lower the value of the thermodynamic quality the bigger the confidence intervals are.

The plots from 5.21 show the tensor smooth functions fitted, in which grey areas indicate regions that are not recommended to make predictions due to the lack of data in this area. Another characteristic of this type of contour plot is presented by the colors, in which the yellow represents higher values of the output smoothing functions, and the blue color is associated with lower output values. For example, values of pressure below 5,000  $kPa$  and mass flux higher than  $20,000 \text{ kg m}^{-2} \text{ s}^{-1}$  will influence the tensor smooth function of this pair to have higher values and make the predicted quantile of the CHF being higher, while mass fluxes below  $20,000 \text{ kg m}^{-2} \text{ s}^{-1}$  will lead to lower values of the function.

Figure 5.22 shows for each observation from the modeling dataset their respective predicted quantile  $\tau = 0.95$  represented by the red dots or line in the case of the plot in the bottom, in which observations that surpass the red limit have a probability of occurrence lower than 5%, or depending on the use case a high-risk observation. This high-risk measurement can affect a CHF control process, like those in nuclear power plants. This type of control can help by adding an operating limit control based on the behavior of the CHF. Besides not being studied in this work, other quantiles can be fitted to understand and predict CHF in different situations where a GAM model is incapable due to its limitations.

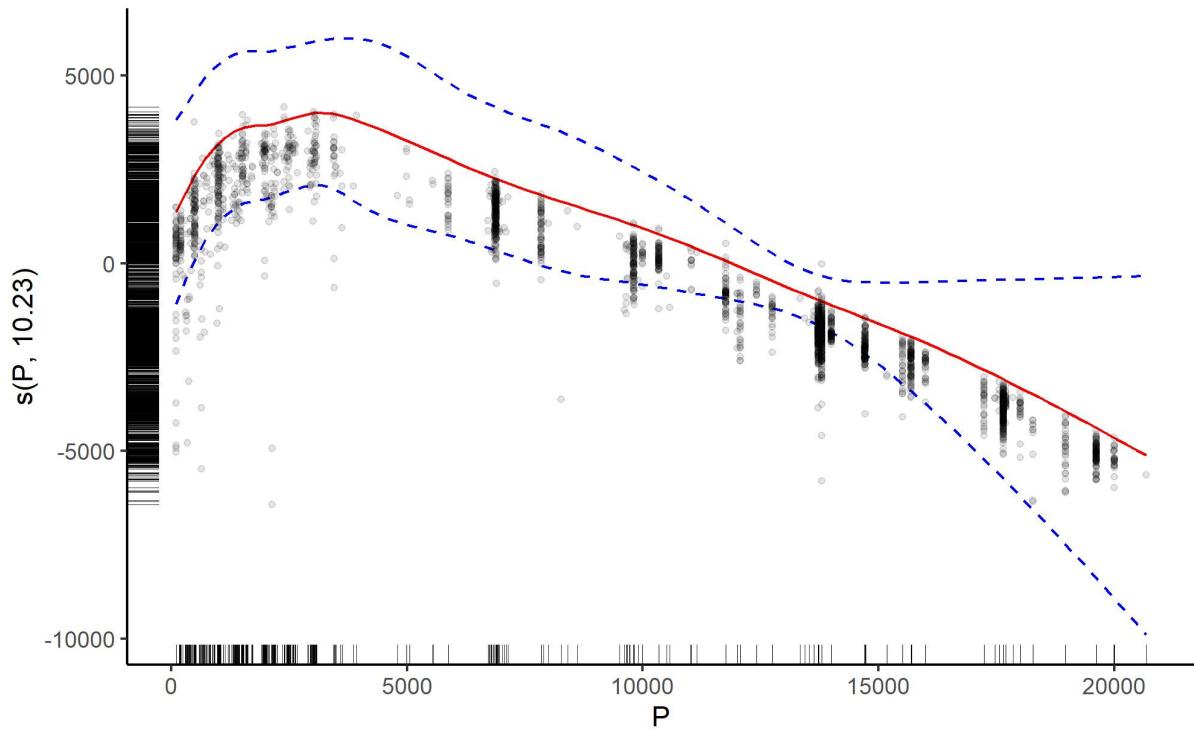


FIGURE 5.16 – One-dimensional smoothing functions for the pressure ( $s("P")$ ) fitted by QGAM model for the quantile  $\tau = 0.95$ .

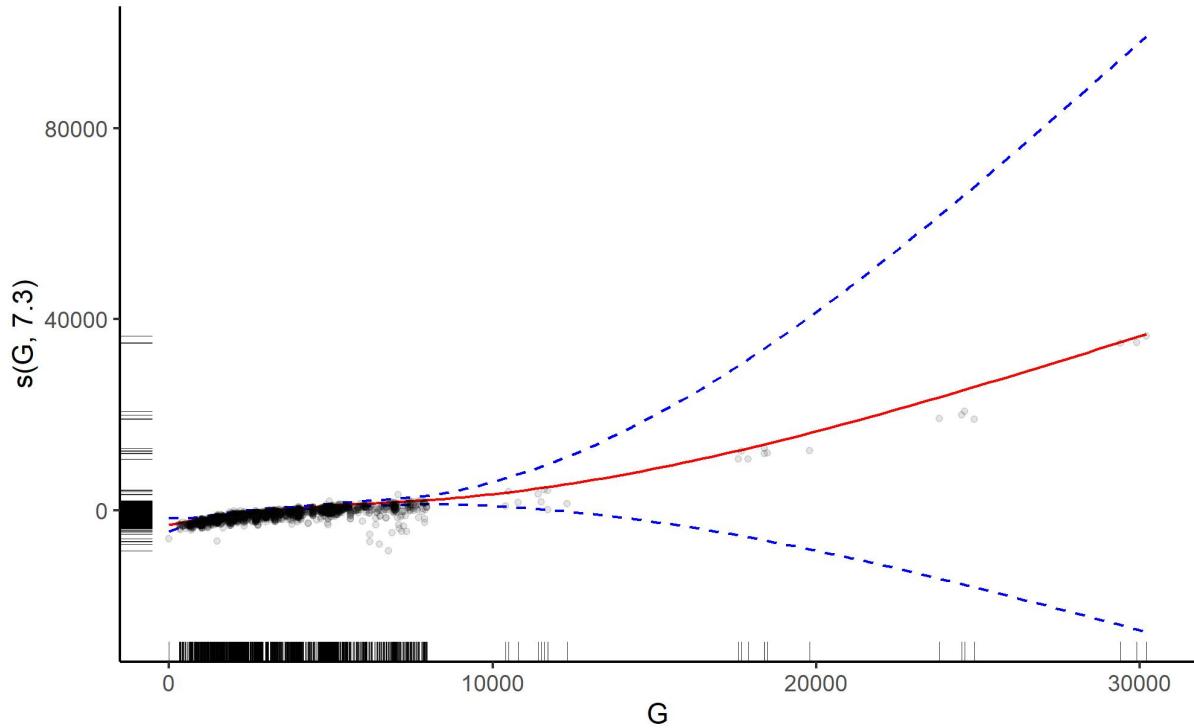


FIGURE 5.17 – One-dimensional smoothing functions for the mass flux ( $s("G")$ ) fitted by QGAM model for the quantile  $\tau = 0.95$ .

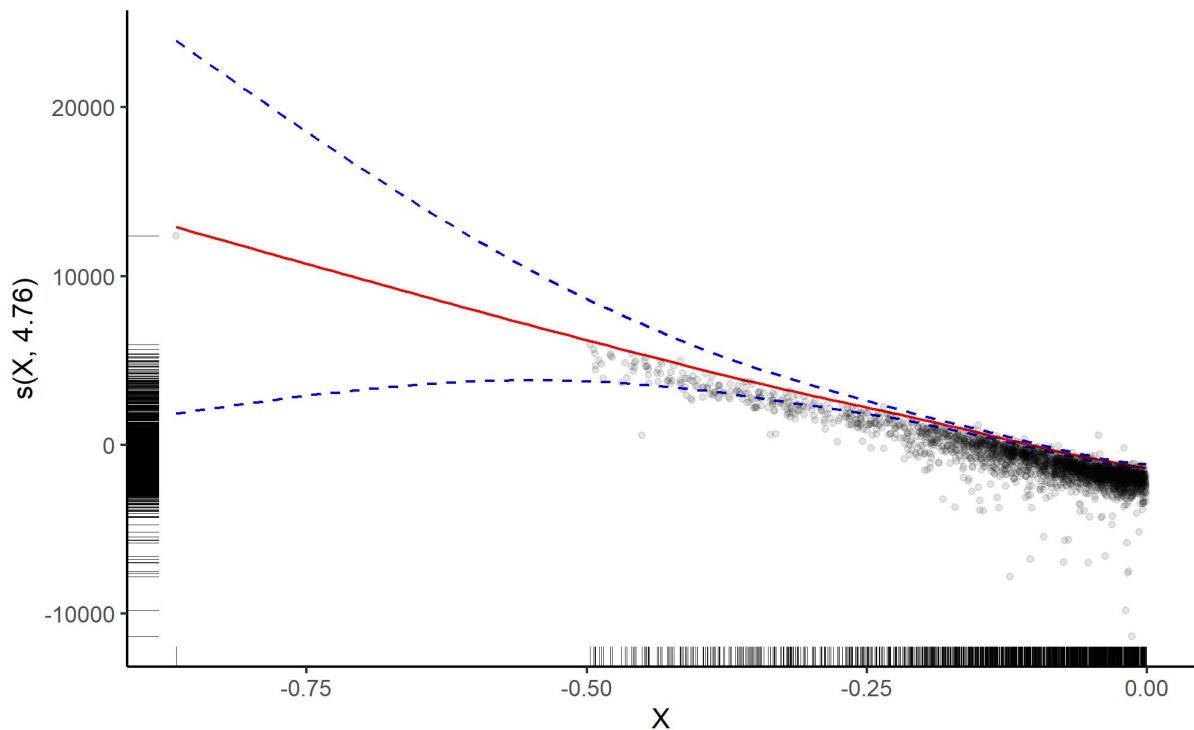


FIGURE 5.18 – One-dimensional smoothing functions for the thermodynamic quality ( $s("X")$ ) fitted by QGAM model for the quantile  $\tau = 0.95$ .

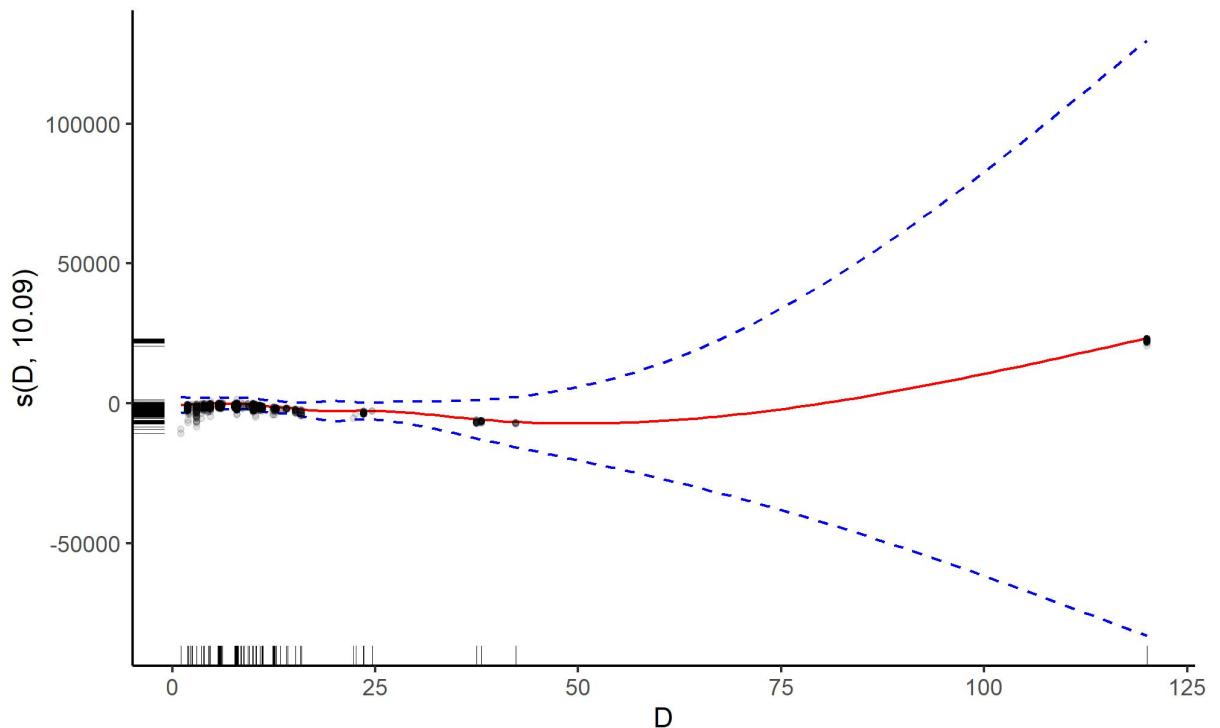


FIGURE 5.19 – One-dimensional smoothing functions for the heated diameter ( $s("D")$ ) fitted by QGAM model for the quantile  $\tau = 0.95$ .

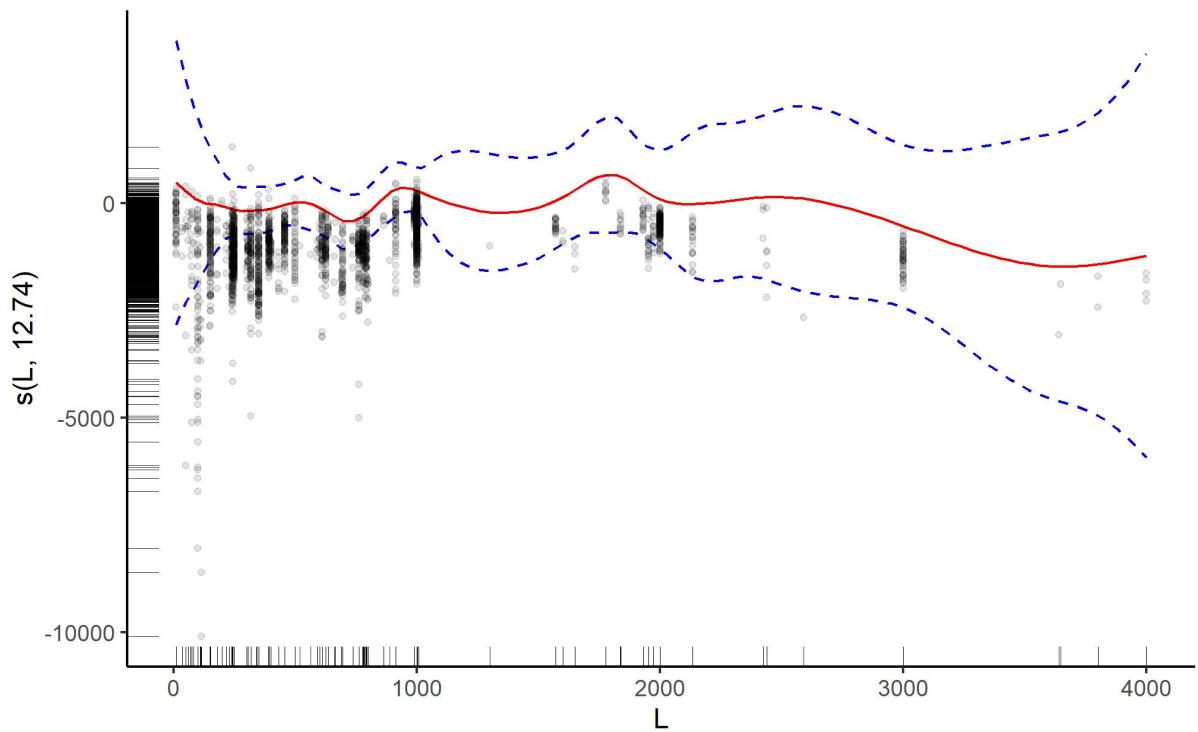
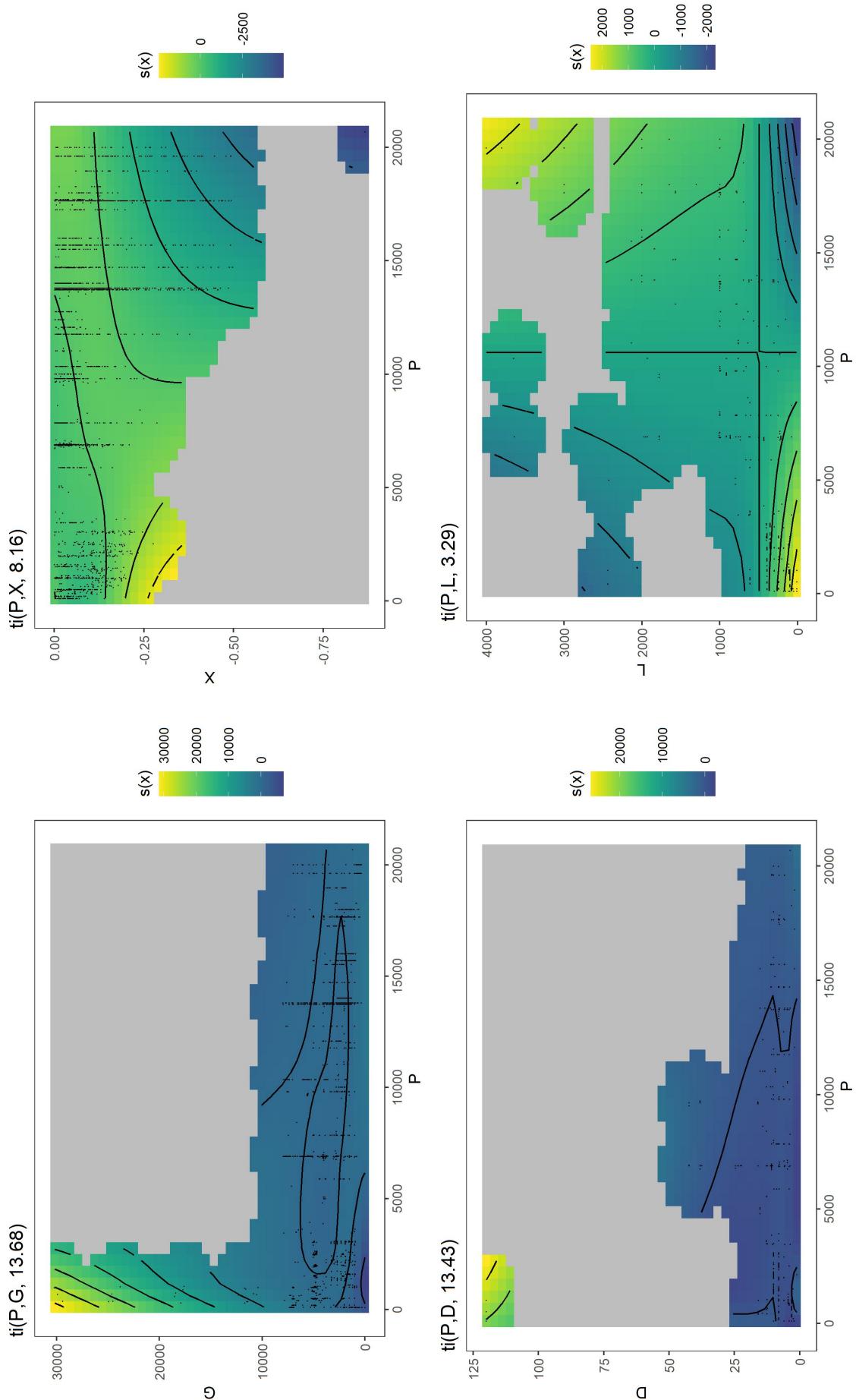


FIGURE 5.20 – One-dimensional smoothing functions for the heated length ( $s(L)$ ) fitted by QGAM model for the quantile  $\tau = 0.95$ .



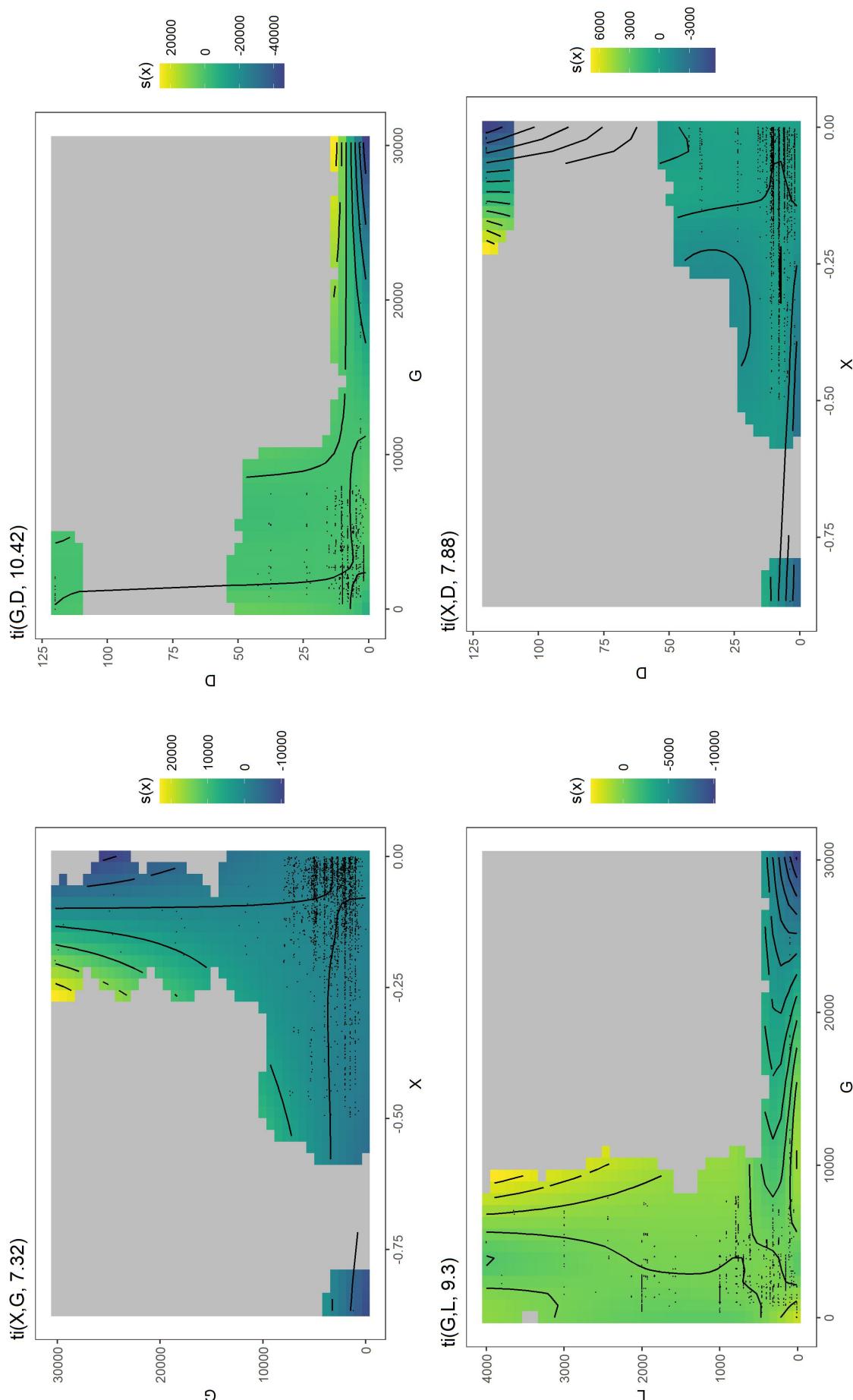
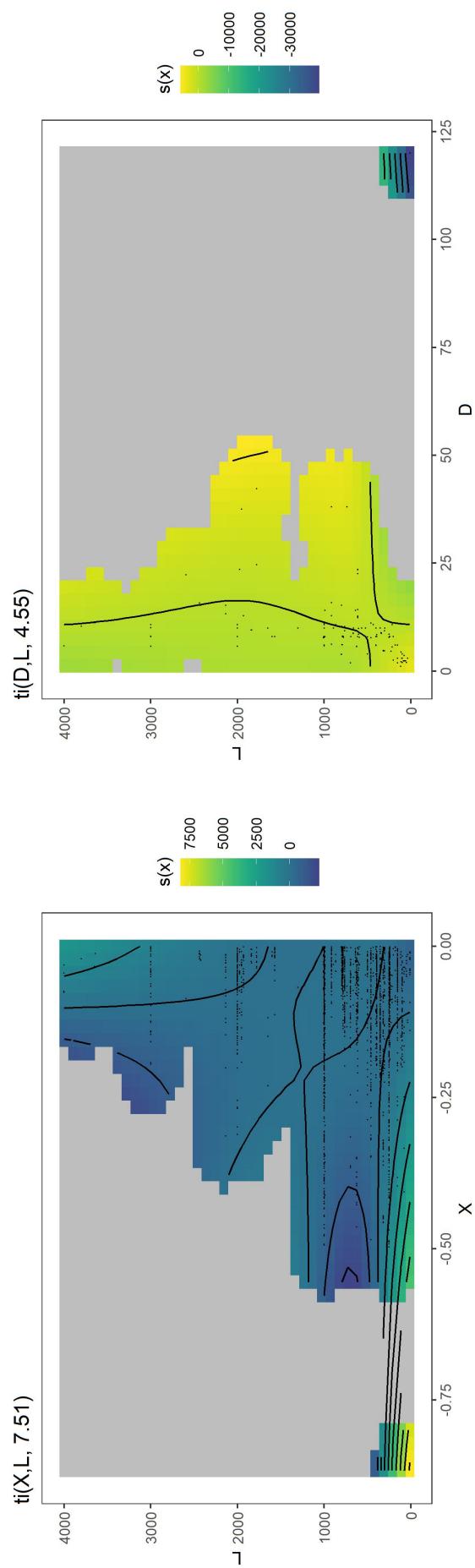


FIGURE 5.21 – Two-dimensional tensor functions ( $t_i()$ ) fitted by the QGAM model for the quantile  $\tau = 0.95$ .



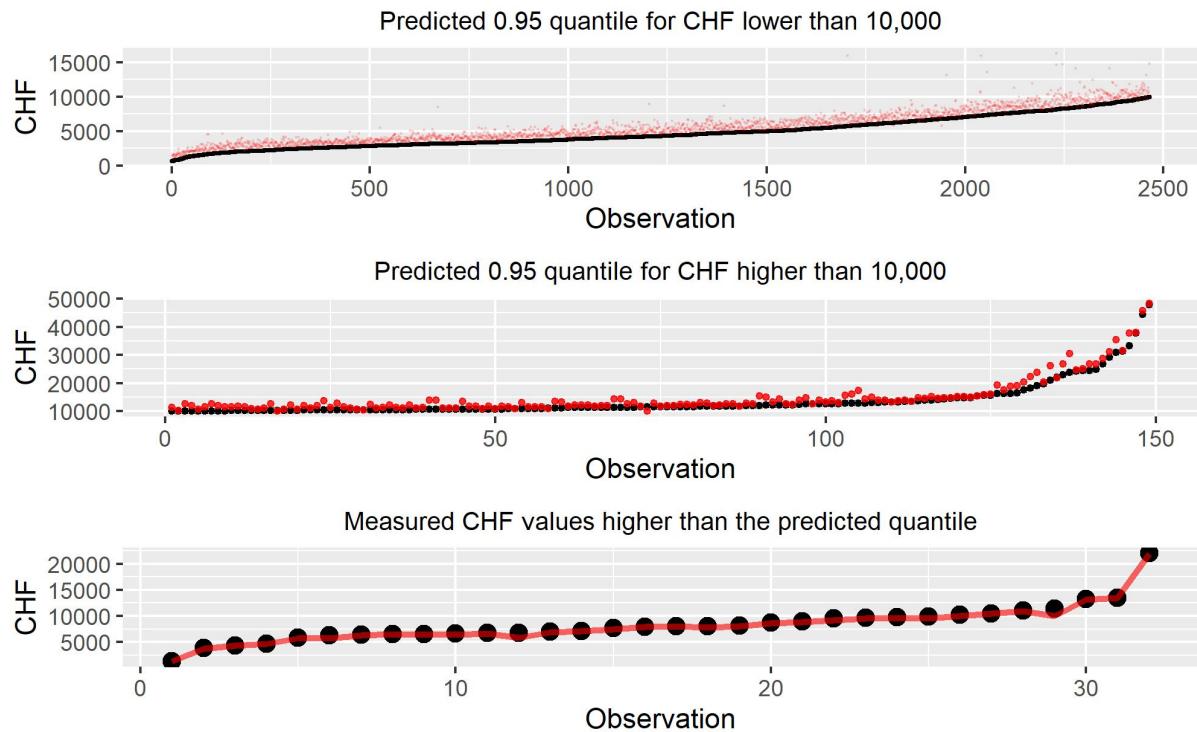


FIGURE 5.22 – Predicted  $\tau = 0.95$  quantile versus measured CHF. The data was ordered by the lower to the higher value of the CHF to facilitate the visualization.

## 6 Conclusion

This work presented four types of statistical additive models: the simple linear regression, the generalized linear regression model, the generalized additive model, and the quantile generalized additive model. These four models were compared with well-established prediction models from the critical heat flux thermodynamic literature: Groeneveld’s lookup table and the Hall-Mudawar correlation model. A compilation of datasets from different sources was collected to train the models and compare their predictive results. The  $k$ -fold cross-validation sampling methodology was implemented to ensure that the fitted additive models do not have their results favored by the sample selected to train the model. We also evaluated the proposed models’ predictive performance on a validation dataset with no observations from Groeneveld’s source, as their dataset has most of the data available in the literature on CHF measurement.

The predictive results obtained in the validation dataset show that the GAM 7 model got the lower  $RMSE_{val}$ , followed by the QGAM 2 model, all the GAMS and QGAMS models fitted outperformed the lookup table and the Hall-Mudawar correlation when comparing their  $RMSE_{val}$ , indicating the capability of the additive models to make predictions of the CHF. An observation-based analysis of the projections from the models agrees with the evaluation metrics and corroborates with the performance of the additive models.

Furthermore, a comprehensive analysis of the best-fitted GAM model was conducted, presenting a range of tools available from a GAM model to aid in understanding the behavior of the fitted model. The visualizations from the univariate and tensor smoothing functions reveal that a simple formula can not fully explain the behavior of the inputs and their impact on CHF prediction. Moreover, these visualizations highlight areas where inference is not recommended due to data scarcity, as their confidence intervals indicate. An analysis of the residuals from the best GAM model, conducted to assess whether the model meets the necessary assumptions, revealed that while the model demonstrates good predictive performance, there is still room for improvement. Adding new data or testing alternative distributions that account for the extreme values associated with CHF measurements could further enhance its accuracy and reliability.

By the end of this work, a separate analysis of the capabilities of quantile additive models to predict extreme values of CHF was conducted, in which a QGAM model for the quantile  $\tau = 0.95$  was fitted and how a researcher can use it in an automated system to control the measured CHF as an upper limit to observed CHF values, reducing the risk in operation systems related to CHF control, as in nuclear power plants.

In conclusion, this dissertation successfully achieved its intended objectives, providing a valuable guide for researchers on the potential application of additive models in predicting subcooled critical heat flux. Additionally, all developed code and datasets used in this work are available at [https://github.com/renands/Subcooled\\_CHF](https://github.com/renands/Subcooled_CHF).

# Bibliography

BARBOSA, R. S.; RIBEIRO, G. B.; SOUZA, C. de. On the prediction of critical heat flux via generalized additive models (gams). In: . **Proceedings** [...]. [S.l.: s.n.], 2023. (International Congress of Mechanical Engineering).

BOX, G. E. P.; COX, D. R. An Analysis of Transformations. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 26, n. 2, p. 211–243, 12 2018. ISSN 0035-9246. Available at: <https://doi.org/10.1111/j.2517-6161.1964.tb00553.x>.

BREIMAN, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). **Statistical Science**, Institute of Mathematical Statistics, v. 16, n. 3, p. 199 – 231, 2001. Available at: <https://doi.org/10.1214/ss/1009213726>.

CAIRA, M.; CARUSO, G.; NAVIGLIO, A. A correlation to predict chf in subcooled flow boiling. **International Communications in Heat and Mass Transfer**, v. 22, n. 1, p. 35–45, 1995. ISSN 0735-1933. Available at: <https://www.sciencedirect.com/science/article/pii/073519339400050U>.

CASELLA, G.; BERGER, R. L. **Statistical Inference**. 2nd.. ed. [S.l.]: Duxbury, 2002. ISBN 0-534-24312-6.

CELATA, G.; CUMO, M.; MARIANI, A. Burnout in highly subcooled water flow boiling in small diameter tubes. **International Journal of Heat and Mass Transfer**, v. 36, n. 5, p. 1269–1285, 1993. ISSN 0017-9310. Available at: <https://www.sciencedirect.com/science/article/pii/S0017931005800961>.

CELATA, G.; CUMO, M.; MARIANI, A. The effect of the tube diameter on the critical heat flux in subcooled flow boiling. **International Journal of Heat and Mass Transfer**, v. 39, n. 8, p. 1755–1757, 1996. ISSN 0017-9310. Available at: <https://www.sciencedirect.com/science/article/pii/0017931095002642>.

CHU, H.; JI, T.; YU, X.; LIU, Z.; RUI, Z.; XU, N. Advances in the application of machine learning to boiling heat transfer: A review. **International Journal of Heat and Fluid Flow**, v. 108, p. 109477, 2024. ISSN 0142-727X. Available at: <https://www.sciencedirect.com/science/article/pii/S0142727X24002029>.

DEMETRIO, C. G. B. **Modelos Lineares Generalizados em Experimentação Agronomica**. [S.l.: s.n.], 2002. Class Notes. Available at: <https://docs.ufpr.br/~niveam/micro%20da%20sala/bom/Apostila%20de%20MLG.pdf>.

- FASIOLO, M.; WOOD, S. N.; ZAFFRAN, M.; NEDELLEC, R.; GOUDE, Y. Fast calibrated additive quantile regression. **Journal of the American Statistical Association**, Taylor & Francis, v. 116, n. 535, p. 1402–1412, 2021. Available at: <https://doi.org/10.1080/01621459.2020.1725521>.
- FASIOLO, M.; WOOD, S. N.; ZAFFRAN, M.; NEDELLEC, R.; GOUDE, Y. qgam: Bayesian nonparametric quantile regression modeling in r. **Journal of Statistical Software**, v. 100, n. 9, p. 1–31, 2021. Available at: <https://www.jstatsoft.org/index.php/jss/article/view/v100i09>.
- GROENEVELD, D. **Critical Heat FLux Data Used to Generate the 2006 Groeneveld Lookup Tables**. [S.l.], 2019. Available at: <http://www-sop.inria.fr/science/skd.gz> http://www-sop.inria.fr/science/skd.gz. Accessed on: 09 Mar 2023.
- GROENEVELD, D.; LEUNG, L.; KIRILLOV, P.; BOBKOV, V.; SMOGALEV, I.; VINOGRADOV, V.; HUANG, X.; ROYER, E. The 1995 look-up table for critical heat flux in tubes. **Nuclear Engineering and Design**, v. 163, n. 1, p. 1–23, 1996. ISSN 0029-5493. Available at: <https://www.sciencedirect.com/science/article/pii/0029549395011544>.
- GROENEVELD, D.; SHAN, J.; VASIĆ, A.; LEUNG, L.; DURMAYAZ, A.; YANG, J.; CHENG, S.; TANASE, A. The 2006 chf look-up table. **Nuclear Engineering and Design**, v. 237, n. 15, p. 1909–1922, nURETH-11, 2007. ISSN 0029-5493. Available at: <https://www.sciencedirect.com/science/article/pii/S0029549307002002>.
- GROSFILLEY, E. H. **Investigation of Machine Learning Regression Techniques to Predict Critical Heat Flux**. 39 p. Dissertation (Master Programme in Engineering Physics) — Uppsala University, 2022. Available at: <https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1678090&dswid=1468>. Accessed on: 29 Apr. 2023.
- HALL, D. D. **Critical Heat Flux in Subcooled Flow Boiling**. 1999. 538 p. Thesis (Doctor of Philosophy) — Purdue University, United States of America, 1999.
- HALL, D. D.; MUDAWAR, I. Critical heat flux (chf) for water flow in tubes—i. compilation and assessment of world chf data. **International Journal of Heat and Mass Transfer**, v. 43, n. 14, p. 2573–2604, 2000. ISSN 0017-9310. Available at: <https://www.sciencedirect.com/science/article/pii/S001793109900191X>.
- HALL, D. D.; MUDAWAR, I. Critical heat flux (chf) for water flow in tubes—ii.: Subcooled chf correlations. **International Journal of Heat and Mass Transfer**, v. 43, n. 14, p. 2605–2640, 2000. ISSN 0017-9310. Available at: <https://www.sciencedirect.com/science/article/pii/S0017931099001921>.
- HASTIE, T.; TIBSHIRANI, R. Generalized Additive Models. **Statistical Science**, Institute of Mathematical Statistics, v. 1, n. 3, p. 297–310, 1986. Available at: <https://doi.org/10.1214/ss/1177013604>.
- INASAKA, F.; NARIAI, H. Critical heat flux of subcooled flow boiling for water in uniformly heated straight tubes. **Fusion Engineering and Design**, v. 19, n. 4, p. 329–337,

1992. ISSN 0920-3796. Available at:  
<https://www.sciencedirect.com/science/article/pii/092037969290007Q>.
- IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina: uma abordagem estatística.** [S.l.: s.n.], 2020. ISBN 978-65-00-02410-4.
- JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R.; TAYLOR, J. **An Introduction to Statistical Learning: with Applications in Python.** 1st. ed. [S.l.]: Springer, 2023. ISBN 3031387465.
- A Brief Review of Computational Intelligence Techniques for Critical Heat Flux Prediction**, Volume 6B: Thermal-Hydraulics and Safety Analyses de **International Conference on Nuclear Engineering**, (International Conference on Nuclear Engineering, Volume 6B: Thermal-Hydraulics and Safety Analyses). V06BT08A050 p. Available at:  
<https://doi.org/10.1115/ICONE26-82325>.
- KALIMULLAH, M.; FELDMAN, E. E.; OLSON, A. P.; DIONNE, B.; STEVENS, J. G.; MATOS, J. An evaluation of subcooled chf correlations and databases for research reactors operating at 1 to 50 bar pressure. In: . **Proceedings** [...]. [S.l.: s.n.], 2012. (International Meeting on Reduced Enrichment for Research and Test Reactors). Available at: <https://api.semanticscholar.org/CorpusID:131764827>.
- KHALID, R. Z.; ULLAH, A.; KHAN, A.; KHAN, A.; INAYAT, M. H. Comparison of standalone and hybrid machine learning models for prediction of critical heat flux in vertical tubes. **Energies**, v. 16, n. 7, 2023. ISSN 1996-1073. Available at:  
<https://www.mdpi.com/1996-1073/16/7/3182>.
- KNOEBEL, D.; HARRIS, S.; CRAIN, B.; BIDERMAN, R. M. **Forced-convection subcooled critical heat flux. D<sub>2</sub>O and H<sub>2</sub>O coolant with aluminum and stainless steel heaters.** [S.l.], 1 1973. Available at: <https://www.osti.gov/biblio/4523826>.
- KWON, Y. M.; CHANG, S. H. Evaluation of very high chf for subcooled flow boiling using an artificial neural network and mechanistic models. In: . **Proceedings** [...]. [S.l.: s.n.], 2000. (Proceedings of the Korean Nuclear Society Autumn Meeting).
- MCCULLAGH, P.; NELDER, J. A. **Generalized Linear Models.** 2nd.. ed. [S.l.]: Taylor Francis Group, 1989. ISBN 9780203753736.
- MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis.** 5th.. ed. [S.l.]: John Wiley Sons Inc, 2012. ISBN 0470542810.
- NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society. Series A (General)**, [Royal Statistical Society, Wiley], v. 135, n. 3, p. 370–384, 1972. ISSN 00359238. Available at:  
<http://www.jstor.org/stable/2344614>  
<http://www.jstor.org/stable/2344614>.
- RASHIDI, M.; Alhuyi Nazari, M.; HARLEY, C.; MOMONIAT, E.; MAHARIQ, I.; ALI, N. Applications of machine learning methods for boiling modeling and prediction: A comprehensive review. **Chemical Thermodynamics and Thermal Analysis**, v. 8, p. 100081, 2022. ISSN 2667-3126. Available at:  
<https://www.sciencedirect.com/science/article/pii/S2667312622000475>.

- TANASE, A.; CHENG, S.; GROENEVELD, D.; SHAN, J. Diameter effect on critical heat flux. **Nuclear Engineering and Design**, v. 239, n. 2, p. 289–294, 2009. ISSN 0029-5493. Available at: <https://www.sciencedirect.com/science/article/pii/S0029549308005347>.
- VANDERVORT, C.; BERGLES, A.; JENSEN, M. An experimental study of critical heat flux in very high heat flux subcooled boiling. **International Journal of Heat and Mass Transfer**, v. 37, p. 161–173, 1994. ISSN 0017-9310. Available at: <https://www.sciencedirect.com/science/article/pii/0017931094900191>.
- WOOD, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. **Journal of the Royal Statistical Society (B)**, v. 73, n. 1, p. 3–36, 2011.
- WOOD, S. N. **Generalized Additive Models**: An introduction with r. 2nd. ed. New York: Chapman and Hall/CRC, 2017.
- ZHAO, X.; SALKO, R. K.; SHIRVAN, K. Improved departure from nucleate boiling prediction in rod bundles using a physics-informed machine learning-aided framework. **Nuclear Engineering and Design**, v. 374, p. 111084, 2021. ISSN 0029-5493. Available at: <https://www.sciencedirect.com/science/article/pii/S0029549321000364>.
- ZUBAIR, R.; ULLAH, A.; KHAN, A.; INAYAT, M. H. Critical heat flux prediction for safety analysis of nuclear reactors using machine learning. **2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST)**, p. 314–318, 2022. Available at: <https://api.semanticscholar.org/CorpusID:255267757>.

# **Appendix A - Complementary Information of the used Datasets**

## **A.1 Distribution Plots of the Used Datasets**

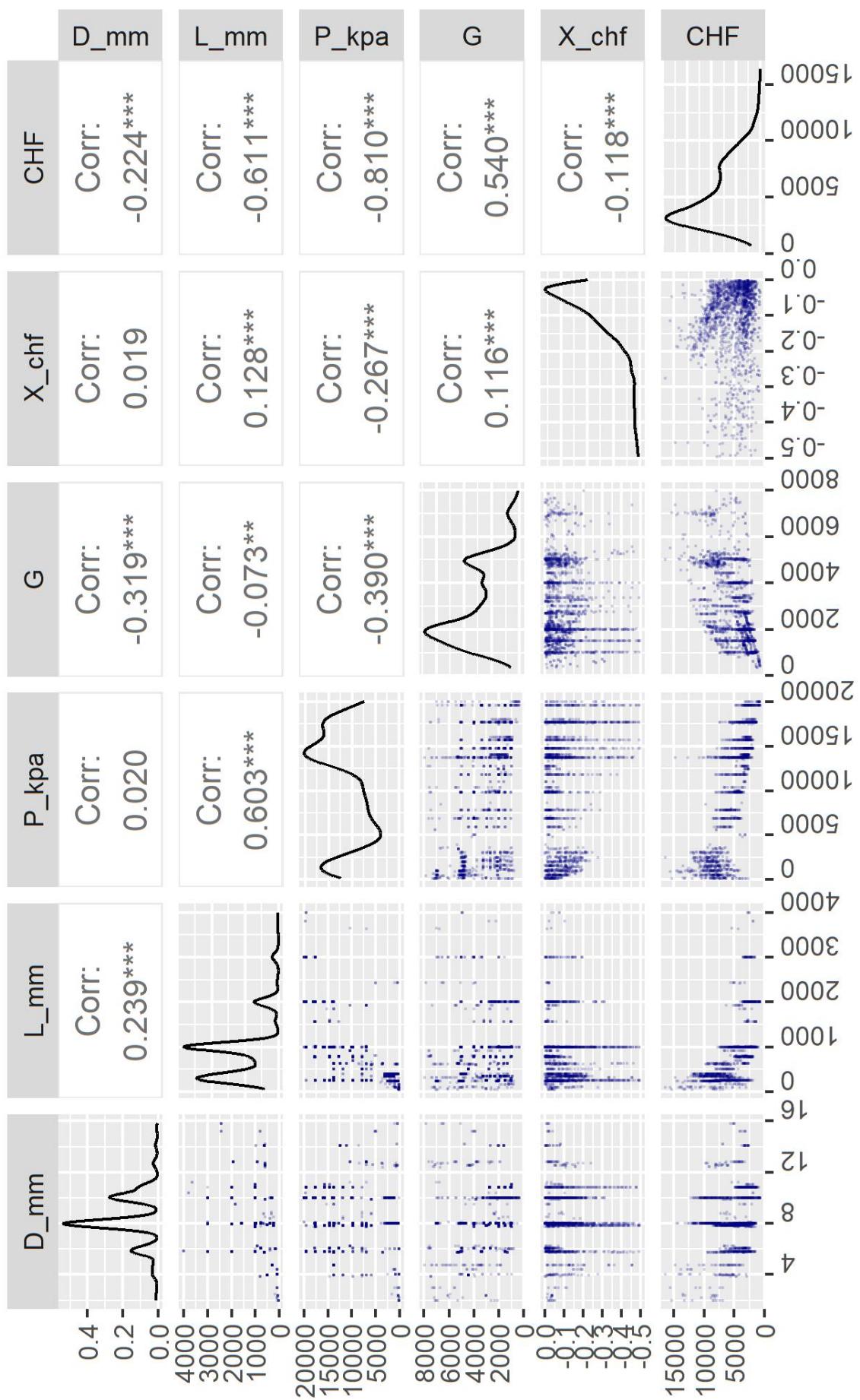


FIGURE A.1 – Distribution of Groeneweld's (GROENEVELD, 2019) dataset, on the lower diagonal, a scatterplot of two variables is shown, the diagonal shows the density plot of one variable, and the upper diagonal presents Pearson's correlation between two variables.

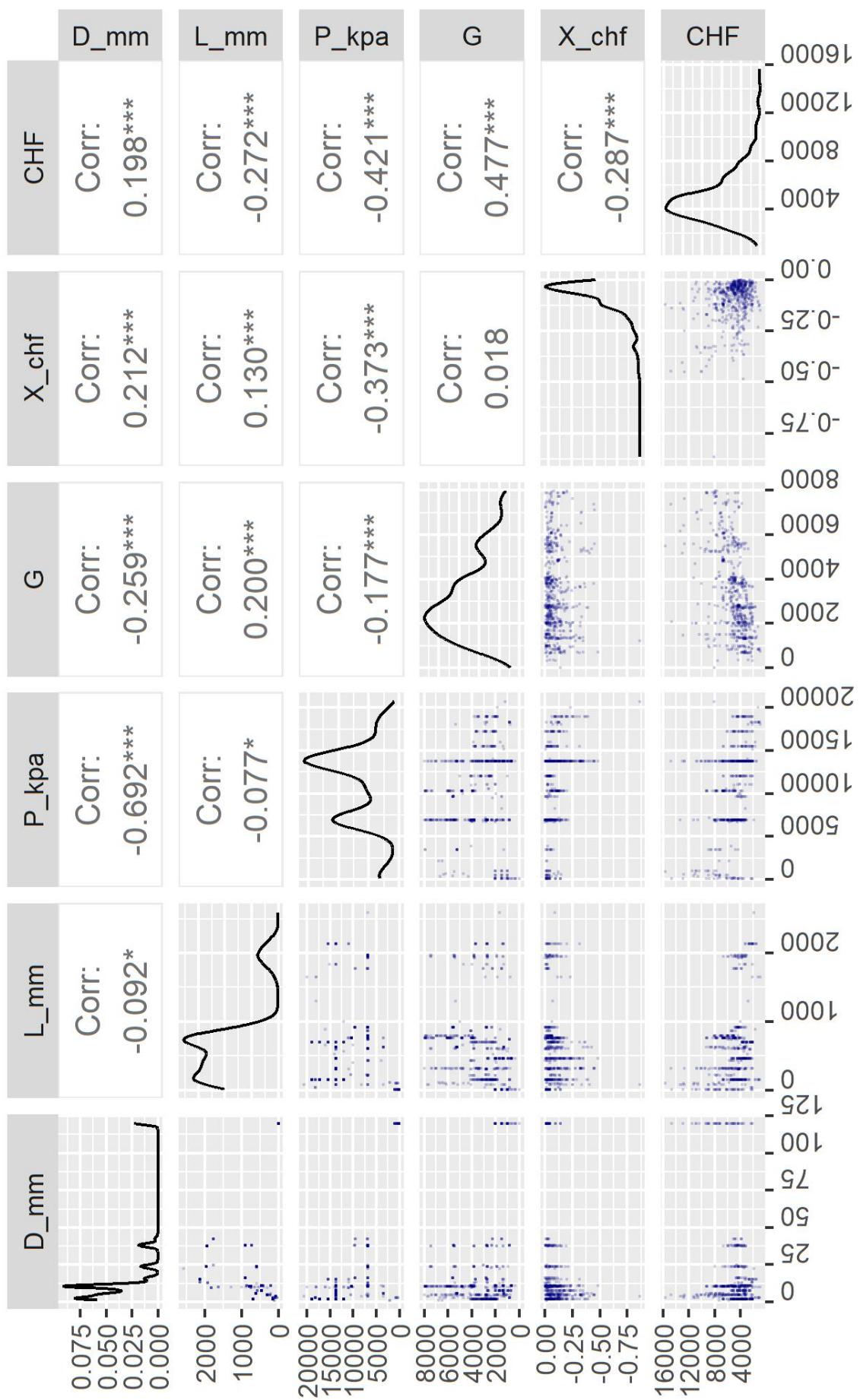


FIGURE A.2 – Distribution of Zhao's (ZHAO *et al.*, 2021) dataset, on the lower diagonal, a scatterplot of two variables is shown, the diagonal shows the density plot of one variable, and the upper diagonal presents Pearson's correlation between two variables.



FIGURE A.3 – Distribution of Inasaki's (INASAKA; NARIAI, 1992) dataset, on the lower diagonal, a scatterplot of two variables is shown, the diagonal shows the density plot of one variable, and the upper diagonal presents Pearson's correlation between two variables.

FOLHA DE REGISTRO DO DOCUMENTO			
1. CLASSIFICAÇÃO/TIPO DM	2. DATA 25 de março de 2015	3. DOCUMENTO Nº DCTA/ITA/DM-018/2015	4. Nº DE PÁGINAS 80
5. TÍTULO E SUBTÍTULO: Additive Models to Predict Critical Heat Flux in Subcooled Flow Boiling			
6. AUTOR(ES): <b>Renan Santos Barbosa</b>			
7. INSTITUIÇÃO(ÓES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÓES): Instituto Tecnológico de Aeronáutica – ITA			
8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR: CHF; Prediction; ; Lookup Table; GAM; QGAM			
9. PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO: CHF; Prediction; ; Lookup Table; GAM; QGAM			
10. APRESENTAÇÃO: <input checked="" type="checkbox"/> Nacional <input type="checkbox"/> Internacional ITA, São José dos Campos. Curso de Mestrado. Programa de Pós-Graduação em Ciências e Tecnologias Espaciais. Área de Propulsão Nuclear e Hipersônica. Orientador: Prof. Dr. Guilherme Borges Ribeiro. Coorientadora: Profª. Drª. Camila Pedroso Estevam de Souza. Defesa em 05/03/2015. Publicada em 25/03/2015.			
11. RESUMO: The critical heat flux (CHF) is the heat flux value in the boiling or cooling process in which the heat transfer decreases, and the heated surface temperature rises rapidly due to factors such as vapor films and bubbles. Due to the difficulties in elaborate experiments and disagreements about measurement and evaluation techniques related to CHF, some methods, like lookup tables, physical correlations, and machine learning, aim to predict the CHF. Additive models are regression models that use an additive structure to define the relationship between the input variables and the CHF to make predictions. This work will compare multiple additive models with consolidated predictive methods in critical heat flux prediction literature. In addition, methods to evaluate and interpret additive models will be shown to demonstrate that additive models can be used as an alternative tool to predict and infer CHF. Furthermore, quantilic additive models will be shown to study the prediction of extreme values of the CHF.			
12. GRAU DE SIGILO: <input checked="" type="checkbox"/> OSTENSIVO <input type="checkbox"/> RESERVADO <input type="checkbox"/> SECRETO			