POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Combined grey- and black-box models with problem-tailored authority distribution

Tesi di Laurea Magistrale in
Automation and Control Engineering
Ingegneria dell'Automazione

Author: **Davide Castagna**

Student ID: 948186
Advisor: Prof. Alberto Leva
Co-advisors: Prof. Simone Formentin
Academic Year: 2020-2021

Abstract

This thesis deals with the combined employment of grey-box models, based on physical considerations or related to the structure of a controller to be calibrated, alongside black-box models (i.e. machine-learning models). The aim is to obtain a higher accuracy of the overall model, consisting of the parallel of the two models, to the data. The distinctive feature of this work is that, unlike most of the literature, the identification of the two structures, which we will conventionally call grey-box and black-box from now on, is not carried out jointly but in cascade. This means that, firstly, the full potential of the grey-box model is exploited given the available data and the purpose of the model, and secondly, a black-box model is adopted for the replication of the residual system.

The motivations for the thesis work are basically twofold. On the one hand, it is required that the authority is distributed between the grey-box model and the black-box model in a way that is best suited to the available information, which is by no means guaranteed if a joint identification approach is used. On the other hand, we want to be able to choose the structure of the grey-box model according to the interpretation we want to give it: this may be physical or linked to the tuning of a regulator. In the first scenario, since the purpose is a simulation, the parameters that constitute the grey-box are quantities with a physical meaning such as resistances, capacitance, etc., whereas if the model identified is oriented towards the tuning of a controller, the parameters can be interpreted as the dominant time constant, gain and delay.

In the course of the thesis work, we will examine some case studies in order to highlight the advantages of both approaches, which are respectively the ability to describe the data as accurately as possible and the ability to physically interpret the results obtained.

Downstream of this analysis, we have actually found that the hybrid approach with variable authority is promising inasmuch, by introducing physical/systemic considerations in the choice of the grey-box model and undertaking the identification of the two models in cascade, we obtain a grey-box model suitable for its purpose (simulation or calibration) coupled with a black-box model that improves its accuracy (beneficial for the validation of the tuned controller).

Keywords: identification, black-box modelling, grey-box modelling

Abstract in lingua italiana

Questa tesi tratta dell'uso combinato di modelli grey-box, basati su considerazioni fisiche o legate alla struttura di un controllore da tarare, insieme a modelli black-box (ivi machine-learning model). Lo scopo è di ottenere una maggiore fedeltà del modello complessivo, costituito dal parallelo dei due modelli, ai dati. La caratteristica peculiare di questo lavoro è che, a differenza della maggior parte della letteratura, l'identificazione delle due strutture, che convenzionalmente da ora in avanti chiameremo grey-box e black box, non viene fatta congiuntamente ma in cascata. Ovvero, si sfrutta, in primo luogo, tutto il potenziale che il modello grey-box è in grado di offrire stante i dati disponibili e la finalità del modello, e in secondo luogo si adopera un modello black-box per la replicazione del sistema residuo.

Le motivazioni del lavoro di tesi sono fondamentalmente le seguenti due. Da un lato, si richiede che l'autorità si ripartisca tra il modello grey-box e il modello black-box in modo al meglio confacente alla informazione disponibile, cosa che non è affatto garantita se si usufruisce di un approccio di identificazione congiunto. Dall'altro lato, si vuole poter sciegliere la struttura del modello grey-box a seconda dell'interpretazione che se ne vuole dare: essa può essere di carattere fisico oppure legata alla taratura di un regolatore. Nel primo scenario, essendo lo scopo simulativo, i parametri che costituiscono il grey-box sono grandezze con un significato fisico come resistenze,capacità, ecc, invece se il modello identificato è orientato alla taratura di un controllore, i parametri sono intepretabili come costante di tempo dominante, guadagno e ritardo.

Nel corso del lavoro di tesi prenderemo in considerazione alcuni casi studio in modo tale da evidenziare i vantaggi di entrambi gli approccio che sono rispettivamente la capacità di descrivere nel modo più accurato possibile i dati e l'abilità di interpretare fisicamente i risultati ottenuti.

A valle di questa analisi abbiamo effettivamente riscontrato che l'approccio ibrido ad autorità variabile è promettente in quanto, l'introduzione di considerazioni fisico/sistemiche nella scelta del modello grey-box e intraprendendo l'identificazione dei due modelli in cascata, si ottiene un modello grey-box adatto al suo scopo (simulazione o taratura) ac-

coppiato a un modello black-box che ne migliora l'accuratezza (proficuo per la validazione del controllore tarato).

Parole chiave: identificazione, modello black-box, modello grey-box

Contents

Introduction

Identifying the reality is a fundamental step that every STEM (science, technology, engineering and mathematics) operator has to deal with. We can represent a real process with a different grade of specificity based on the objective of the identification but we have to always remember that reality is the most complex process physically appreciable. Therefore, the representation of nature can be described as a picture in which the contribution of every pixel is paramount or as a sketch where the structure is enough. The deepness of the image is not set a prior for every type of model but reflect the aim where the model will be used. So, generally, the reality can be reduced to a simplified model that accounts only for the behaviour of some physical variable or we can develop a complete online adaptive model that takes into consideration some technical and operational aspects like:

- time-varying characteristic: there is no particular single operation point around which the identification can be designed;
- nonlinear behaviour: it is not possible to approximate the system to a model linearized around a single operating point;
- model inaccuracies;
- few specific measurements: the available measurements of physical quantities might present low accuracy due to the wide range of operations that the measuring instrument has to cover;
- presence of disturbances: Some disturbances cannot be totally ruled out. These disturbances can be due to operator mistakes, processing problems, presence of an impurity in the product;
- irreversible behaviour.

Anyway, the way how the model describes reality is a matter of what and how much information about its nature is able to capture. Generally, the real system information can be subdivided into three categories. There is information about the knowledge of the system that carries the description of the first principle or its empirical representation adopted in the past. The second category is the data, a data-set is a container of information that

describes the nature of the real system in a different perspective where just a frame of the whole movie is stored. Another carrier of information is the assumptions that establish and contains the relationship between the model and the real system. The calibration between this 3 kind of information, distinguish a variety of model types. To be precise, the assumptions are involved in all the types of models so the real distinction between the representation of a process is achieved by the amount of physical and data information that we introduce.

On one hand, there are models built only using physical information called the white-box model and on the other hand, there are models created by correlation introduced by the data called black-box model.

The adoption of one of the two extremes provides an incomplete description of the real world complexity because a purely physical description of the process leads to a stiff model lacking the ability to counteract parameter uncertainties. On the other side, is improbable that we are dealing with a physically unknown process and at the most the rejection of the physic that governs the real application constraints the data-driven model to a certain range of work conditions. In other words, the generalization property of a model is restored by the adoption of a physical oriented model, and the interpolation quality is supplied by taking into account data.

Then, in some respect, all the models are a combination of physical oriented and data-driven models. An explicative example of this situation is a model created by differential equations with a handful number of the parameters that are unknown and have to be retrieved by a data-based approach. This scenario is called grey-box modelling.

Research Aims

In this thesis, we want to debate an approach that combines grey-box modelling based on the first principles and a black-box data-driven modelling approach with arbitrary techniques.

Data-driven approaches are the most disparate, especially the arrival of machine learning techniques increase even more the available choice because it allows the development of high-performance model able to make very accurate predictions and decisions on a wide range of applications without understanding and explaining the model's prediction mechanism. The simplicity and the accuracy of machine learning models to extract useful and valuable information from large amounts of data is the main advantage with respect to the mechanicistic approach of modelling. However, artificial intelligence and machine

learning trends in the system identification scenario move towards an opaque identification where the effort invested in a deep understanding of the physical behaviour is refused. Paradoxically, blind reliance on data can possibly lead to an unfair model because this enormous amount of data may contain biases. So the possibility to inadvertently identify spurious correlations in the training data exists that can impact on the reliability of the overall model.

Hence, in our point of view, reversing the current trend can only benefit the identification approach otherwise we risk creating and using a model that we do not really employ because its reliability is connected with its capability to work in different scenarios that are which are nevertheless physically acceptable by the underlying physic.

In the literature, there are a lot of examples where physical knowledge and data knowledge are adopted to create a model that is interpretable and at the same time accurate. I quote for illustrative purposes the paper written by Sohlberg B. and Jacobsen E.W. [16] in which they listed the main hybrid approach and defines the main advantages of the abovementioned method. This study is exemplifying because they use as black box part three different types of model, from a Taylor approximation model to an ARMAX model. The conclusion brought by the authors are promising and they justify the adoption of a mixing grey-box and black-box approach as being more effective than a pure black or white box model. What most studies in the field of hybrid identification have in common is the choice of an a priori authority granting more responsibility to one approach (black-box) than to the other (grey-box) and vice versa. The most recurrent choice is generated and almost imposed by the trend described above. In fact, reliance on a data-driven method is increasingly being contemplated without scientific reason. Or rather, the main considerations are based on the difficulty of creating an approximate model of reality and the accuracy that an approximate model can achieve. But as we shall see, the allocation of a priori authority can be limiting. In the following chapters, we will define a methodology where authority is chosen automatically by the system based on the data provided. In this way, we create a model with as few compromises as possible, it is able to be physically interpretable and equally accurate because it is capable to find excellent correspondence with data.

The motivation behind this thesis stems from established practice in the literature of prioritising a black-box model over a grey-box model. The assignment of the authority cannot be set a priori, giving an excessive responsibility to a data-driven method, it is very likely that the complete model if subjected to data representative of an operative space different from the acquisition one, will introduce spurious correlations. Conversely, if the authority prioritises a priori a mechanistic model, its interpolation capacity is diminished

even though it will never introduce relationships that are not physically explainable.

The subject of this thesis will therefore be a new interpretation of a methodology already present in the literature. That is a hybrid identification where the authority of the grey-box model is assigned a posteriori. We demand the physics-based model to justify the data with first-principles equations. The accuracy recorded by the grey-box model will be an intrinsic index of the authority assigned to it. Subsequently, the residual generated by the inability of the mechanistic model to fit the data will be identified by a black-box model.

Objectives

The reversal of direction that this thesis seeks to support is justified by four considerations that will be amply validated by case studies in the following chapters. The observations I am going to list can be considered as anticipations of the conclusions and help to define in a deeper way the thesis activity carried out.

1. The first advantage that a variable authority approach can provide is the independence of the method from the grey-box structure. Through the analysis of today's scientific literature, which will be carried out in the next chapter, we note that the usefulness of a data-driven model can take on different orders of utility within a hybrid structure. The heavy influence of a data-only method is required when the first-principles model is simple and therefore not able to approximate the statics and dynamics of a complex real system. Indicatively, if the real process is a first-order linear system, the implementation of a black-box component within the overall model is not evaluated since a trivial grey-box model based on first-order equations is sufficient to mimic the system under analysis. The opposite case is when the identification system is required to recognise a complex system, which attaches non-linearity and interacts in space and time in a difficult way. In this type of scenario the native complexity of these processes induces the identification system to make a choice depending on whether the main interest is. It could be in the development of an interpretable model or the need for an exact model, a simple model or a computationally efficient one, a model robust to uncertainties or with the ability to make future predictions. In these circumstances, the most widely used approach is to place a lot of responsibility on the black box when perhaps a careful analysis of the physical process describing the actual application would lead to a system that can identify parameters that have greater scientific reliability. The methodology we will describe, therefore, does not make assumptions before we even

know and analyse the real process, as we may fall back on an inefficient choice to the exclusion of the other opportunities that systems identification theory provides. The primordial choice of giving more priority to a system with little physical knowledge is in contrast to the aleatory need for a data-driven system that I introduced in a few lines in the previous example. Freeing ourselves from the irrational practice of assigning a priori authority to the black box model allows us to evaluate different grey box structures and verify the advantages and disadvantages that a model more or less related to physics can introduce. In particular, in the following chapters, we will introduce case studies presenting non-linearity and evaluate the adoption of grey-box models that - describe in depth the physics of the real system, e.g. models formed by differential equations and which consider even non-linear dependencies between parameters. - limit themselves to describing only the linear part of the system and do not investigate the complex dependencies that today's physics considers - describe reality in a very approximate way, identifying the elaborated dynamics of reality as a first-order dynamics.

2. The second conclusion corroborates the justification for studying a hybrid methodology based on variable authority. The second justification for the thesis work derives from the previous conclusion since it identifies a further advantage of supporting an approach that is independent of the grey-box structure. The adoption of an extremely simple grey-box model structure is widespread in the systems identification landscape as the availability and easy usability of data benefits the use of methods that do not involve a parameter and reality match. This trend can be positive as the development of an identified model requires little computational power and is able to provide excellent results in terms of accuracy but not all that glitters is gold. The close dependence of the black-box model on the data undermines the correctness of the model because the only ability that the data-driven model is able to provide is the strong interpolation capability of the data submitted to it during the training phase. This means that the black box model is limited in learning the physical relationships in the data because it is constrained by the information that the training data set possesses. Thus, one of the significant critical issues of a model that only identifies relationships in the data is its inability to generalise the identified relationships to a domain greater than that used in the training phase. This is because a black-box model is not motivated by physics. This small introduction helps me to define a problem that might occur if the authority assigned to the black-box is high against a minimal physical description provided by a grey-box model (little authority assigned). In this scenario the contribution of the mechanistic model is

misleading as it introduces a physics that is little detailed so the responsibility of identifying the actual process is delegated almost entirely to the data-driven model. A workload that a flexible structure, such as a black-box model (even better if based on machine learning), is acceptable because it is extremely powerful tool. However, the initial conditions provided by the grey-box system are not complete and force the black-box model to identify the parameters of the system in a environment that is foreign to it (outside its training zone). Given the poor generalisation ability of the data-driven models, the overall model formed by the union of the grey-box model and the black-box model will provide a worse result than the one that the adoption of the grey-box model alone would have provided, since the black-box model introduces spurious correlations to be able to mimic the dynamics of the real system. The implementation of an approach that does not impose a priori authority would have given, in this specific scenario, a higher priority to the grey-box model in order to better identify the real process and avoid the introduction of correlations that are not physically motivated.

3. The third reason why the use of a variable authority is recommended originates from the critical analysis of the current research trend related to system identification. The main flow of resources is channelled into the continuous search for approaches that are disconnected from the reality of the process but extremely efficient. In the previous point we analysed a circumstance in which the data-driven model possesses a high authority and concluded that it is possible to obtain a worse identification in the face of an inaccurate grey-box model. The counter-evidence to this statement is presented in the scenario where the object to be identified is complex and authority is fully vested in a model based on very approximate physics. Therefore, the presence of a black-box model in parallel is to be considered irrelevant as it assumes a priori a negligible authority. In this circumstance, the process of identifying a complex real system with an extremely simple physical model provides a very approximate model of reality. But the model does not introduce spurious correlations, the data represented have a direct correspondence with physical parameters, even if not correctly identified or own a too small multiplicity compared to those required by reality. This thesis is confirmed by Timur Bikmukhametov and Johannes Jäschke [17] in a paper they published in 2020. In particular, they consider various combination solutions between first-principles and machine-learning models, analysing the scenario in which a real case study is subjected initially to a grey-box identification and subsequently to a black-box identification. Interesting are the considerations they make at the conclusion of the first phase, which confirm

the generalisability of a grey box model even if it does not reach a higher accuracy than a well-calibrated hybrid system. In other words, an incorrect distribution of authority in favour of a model based on physical principles does not present worse results than an inverse distribution could do. Because, as I have explained above, it could introduce dynamics not correlated with the reality of the process.

4. The last point underlines the polyvalence of proposed approach. It is not new that a data-driven methodology such as black-box identification is adaptable to multiple problems because its nature is stochastic and not deterministic. In the common scenario, identification using a grey box model is carried out in parallel with system identification using data-driven methods. This circumstance limits the potential that a combined approach can express in terms of generalisability because it lacks the analysis and interpretation of the underlying physics. The introduction of a standard grey-box model which is the same for all processes to be identified limits the generalisation properties of the overall model because the excessive increase in authority of the black box model imposes its interpolation ability in spite of greater generalisability. The approach we describe is multipurpose and suitable for a greater variety of applications because it is independent of the structure of the grey box model. It can be chosen according to the level of interpretability required or according to the purpose for which the model is designated. If a process is particularly complicated, such as a chemical or biological process, and our aim is to faithfully mimic the dynamics of the real system, the structure of the grey box will be more articulated. It will also consider non-linear dependencies so as to produce a model that is understandable, akin to physics and usable even in circumstances of which it has no experience (different working points). Similarly, if the purpose of identification is to create a model that will be used for the synthesis or calibration of a controller then it is advisable to implement a grey box model of appropriate structure. However, the versatility that such an approach introduces is justified both by the way in which authority is partitioned but also by a study of the physics governing the process and the purpose for which the model is created. The versatility of the method is vanished if there are no interpretable and physically recognisable data because it is tricky to define general rules governing all real processes.

Finally, the final conclusion is a proof of the motivations just listed. The definition of a versatile approach free from the grey box structure favours the use of a model identified for several categories:

- the first category is the analysis of system behaviour;

- the second category is the design and analysis of a system structure for a required operation;
- the third category is the design and analysis of a system controller for a required operation.

The multidisciplinarity that an approach like this provides, arises from the possibility of varying the structure of the grey box model according to the purpose of the identification. If the purpose of identification is as described by the third category, some considerations need to be added to get a clearer picture. Broadly speaking, adopting a simple grey-box model and a complex black-box model may not always be the right alternative if one wants to synthesise a controller. Certainly, the development of a control system will be designed considering only the mechanistic component of the model and will not take into account the empirical part. However, if the data-driven component of the model has a high authority, the controller may perform less well than in the case where the grey box model is only required to model small uncertainties. In other words, the integration of a black box model with high authority may worsen the synthesis of the controller as it may introduce correlations that do not exist in the physical model. Then, the possibility of individually calibrating the authority of one approach with respect to the other allows the overall model to be suitable for synthesising controllers acting on complex processes. This framework strengthens the motivation of the thesis work because it is an approach that can lead both to a higher performance of the controller based on the identified model but at the same time allows to design a controller that is more reliable.

The motivations I have listed will be detailed in the following chapters both on a theoretical level, i.e. by introducing the methodology and comparing it with the know-how present in the literature, and on a practical level by applying the procedure to two specific case studies. The methodology has been applied to two case studies with different characteristics in the time and frequency domains in order to provide a complete picture, also analysing borderline cases in which the implementation of a black-box model with high acquired authority can lead to disadvantages in the replication of a real process as it introduces spurious correlations (not physically explainable). More deeply, we will observe that in the first example (thermodynamic case study) the implementation of a more or less complex grey-box model will be able to identify the main heat transfer dynamics and the addition of an ARX model will allow to mimic the residual dynamics always improving the fit of the overall model with respect to the real system. The variable that will most affect the accuracy of the overall model is the complexity assigned to the grey-box model. In other words, the complexity of a grey-box model used to tune a controller will give a worse contribution to the combined model than a more complex model used to accurately

simulate the case study.

In the second example (electrical case study) the conclusions obtained in the first example are valid but the study of the above application will allow to validate the approach also to a process that shows a different behaviour in the frequency domain compared to the first case study. The applicability of the proposed procedure in two different areas with different characteristics enhances its possible generalisability.

A possible representation of the two real applications can be as follows (Fig.1).

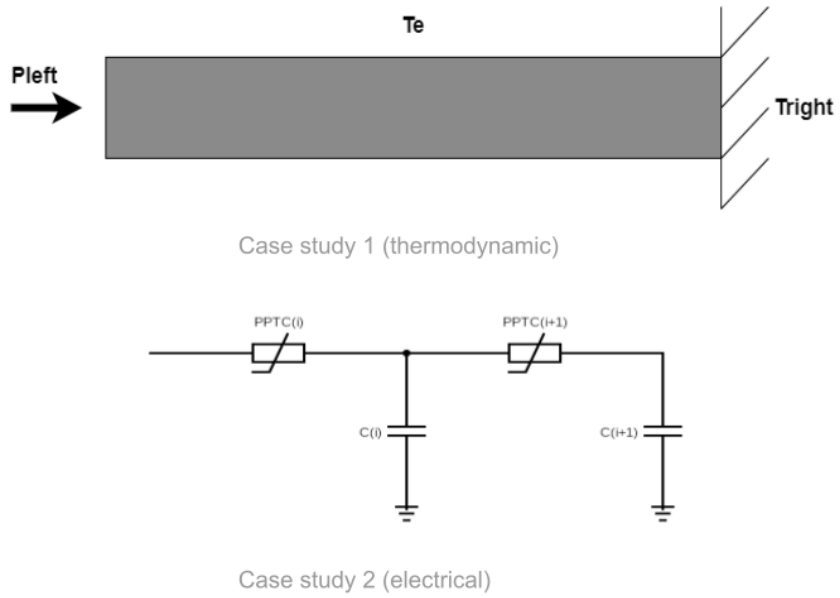

Figure 1: Representation of the two case studies

Outline of the thesis

The thesis is organised as follows.

1. **Chapter 1** analyses the available scientific literature in such a way as to contextualise the thesis research in the overall scenario. In addition, it concerns the state of the art regarding identification with black-box, grey-box and hybrid models.
2. **Chapter 2** presents the proposed methodology, defining a procedure that starts with the analysis of the data, proceeds by illustrating the purpose of the identified model, and ends by presenting the procedure including some considerations.
3. **Chapter 3** applies the procedure outlined in the previous chapter to two significant

case studies. The first concerns the study of heat transfer in a rod. The second involves the analysis of an RC circuit with Varistor resistors, the so-called PPTC (poly switch). These two experiments are relevant because we will study two systems which present non-linearity in time and non-linearity in frequency respectively.

1 | Literature Reviews

The aim of the scientific community to increasingly introduce data-driven models stems from the need to model systems that exhibit inaccuracies and/or non-linear behaviour, are located in disturbed environments, or are difficult to extrapolate physical knowledge because they are too complex. This trend is computationally advantageous because it can produce an extremely accurate result with minimal effort, but at the same time the solution is extremely complex to analyse. As I will describe in the next sections, the interpretability of a model introduces numerous advantages and therefore the use of a model based on first principles should not be demonised a priori but can introduce benefits of a different nature such as physical awareness, generalisability, etc.

1.1. Machine Learning

With the advent of machine learning and the consequential development of high performance machine learning model, we are able to make very accurate predictions and decisions on a wide range of applications without understand and explain the model's prediction mechanism.
The simplicity and the accuracy of machine learning models to extract useful and valuable information from large amounts of data is the main advantage respect to the mechanicist approach of modelling. However this enormous amount of data may contain biases that can possibly lead to an unfair model. So the possibility to inadvertently identify spurious correlations in the training data exists that can impact on reliability, on safety and on industrialiability of the overall model. Therefore, we risk to create and use a model that we do not really understand because we are not able to interpret the underlying rationale. Likewise, the use of machine-learning models, doesn't allow the operator to identify and correct an error that occurs in a specific variable because the correspondence between the parameter and the reality is not contemplated. Furthermore, the tight dependency of the machine learning model to the data prejudicates the correctness of the model itself because to train a machine learning model we need an input signal that make possible the extraction of useful and valuable information. However, it is an hard task to inject a

sufficient labeled data because it is a opaque data-set where unlabeled data easily arises. Therefore, we are obliged to enlarge the initial small labeled data-set from a rather large unlabeled data-set and iteratively extrapolate the most confident predictions. An extra step that augment the accuracy of the model but unfortunately decrease a lot the explainability and the interpretability of the predictor.

1.2. Need of Interpretability and Explainability

The interpretability and the explainability constitute key factors that are of high significance in ML practical models. Interpretability is the ability of understanding and observing a model's mechanism or prediction behaviour depending on its input stimulation while explainability is the ability to demonstrate and explain it in understandable terms to a human. Essentially, an explanation is an "interface" between humans and the machine that is at the same time both accurate and comprehensible. An explanation could be also relevant for two cases: it is fundamental for revealing outcomes in data that explain the identification algorithm of the machine or to explain how the identification itself works. In other words, an explanation is required both for the theoretical research that has to explain the logic behind a data-driven model and for the application user that aim to explain why a certain decision has been returned for a particular input.

The interpretability of the model is taken into consideration with different weights. The first peculiarity aspect is related to how much a model can be interpreted, we can deal with a completely or locally interpretable model. Hence if we are able to understand the whole logic of the model we are speaking about global interpretability, instead, we indicate with local interpretability if it is possible to comprehend just the reasons of a particular prediction. Another aspect that we have to take into account is the nature of user expertise because if the user has huge background knowledge and experience, he is able to interpret a sophisticated model and prefers to handle a more opaque one.

The next step to define is the aspiration that an interpretable model has to accomplish, the purpose. Thus, the introduction of desiderata are essential. Firstly, we need to define how much the model is interpretable. The most addressed way to measure it is about the complexity of the model in terms of model size. Secondly, we have to describe to which extent the model accurately predicts unseen instances. The adoption of the accuracy score, the F1-score can evaluate the grade of accuracy of an interpretable model. Otherwise, we can classify it in terms of fidelity. That is, the ability to faithfully imitate a black-box predictor and how much the outcome of the interpretable model mimics the outcome of the non-interpretable model. Furthermore, machine-learning models should also have other

ordinary important required features such as reliability, robustness, causality, scalability, and generality. This means that a model should have the ability to maintain certain levels of performance independently from small variations of the parameters or of the input data (reliability/robustness) and that controlled changes in the input due to a perturbation affect the model behaviour (causality). Moreover, it is appropriate that the model is not constrained to certain particular initial conditions or training restrictions but the model must be usable on a large scale and in a different framework(scalability/generality).

In the state of the art, a small set of existing interpretable models is recognized: decision tree, rules, linear models. These models are considered easily understandable and interpretable for humans.

In short, decision trees and rules classifiers are models that are governed by a set of decision laws with the if-then form where the outcome is associated with a label class. Thus, the unlabelled data that is classified to an opportune labelled class embeds a kind of clue about the classifier itself. In this way, the extrapolation of a part of the knowledge that governs the model is straightforward. However, the main approach that I want to examine are linear models. An explanation of a linear model can be done considering the feature importance. If some value, that intrinsically represents an attribute, contributes substantially to the model's output, then it means that the corresponding feature has an higher influence on the prediction of the model. However, the evaluation of the contribution of the main features become harder when the model doesn't fit the training data and it is forced to introduce spurious correlation to optimize the error between the prediction of the model and expected predictions. Moreover, the recognization of the main features is impossible when the size and the complexity of the linear model increases because the classification algorithm is humanly unmanageable.[12]

1.3. From data-driven to first principle model

The last section introduces the concept of interpretability and tries to match it with the world of machine learning techniques. However, it is clear that is an extremely hard task, especially when the model is complicated because machine learning is stochastic, not deterministic. The aim of this type of approach is to accurately mimic the pattern inside the data, therefore forcing the machine learning algorithm to provide an explanation of the underlying physics is a stretch [8]. While it is plausible that information about causal processes will ultimately prove relevant to knowing the reasons for outcomes, talk of interpretability in machine learning is misleading insofar as it presupposes that an interpretation will satisfy the grade of physical details that govern the process. For this

reason, the state of art suggests using machine learning techniques when the physic is too complex to describe because, in these particular circumstances, the ability to find complex patterns without providing an explicit form of them is the unique available solution. Moreover, the lacking of transparency makes the data-driven models easier to construct in comparison to first principles models.

To sum up, the added value that the machine learning approach provides to the model identification is confined to the capacity of achieving a high level of accuracy in the results. We will prove that if the system behaviour is relatively complex and the constructed first principles models are relatively simple, it is better to use a pure data-driven model, a simple yet accurate solution. This is because, for complex system behaviour, simple first principles models may not be accurate enough to accurately represent it. At the same time, informativeness in data may be well-correlated with the target process, such that the model produces accurate results. When the system behaviour has a moderate complexity, so that even simple first principles models are accurate enough, its combination with the raw data may be a good choice. Consequently, the deterioration of the performance of a model composed of two entities, the mechanistic one and the data-driven one, arises when the description of the physic through ordinary differential equations is too scaled-down compared to the complexity of the real process. The introduction of the first principle in the overall model is a potential advantage for multiple factors:

- brings interpretability and explainability;
- increases the accuracy if the model describes the most physical meaningful behaviours;
- improve model generalizability, so perform well on the unseen data;
- able to describe dynamics that have been barely seen in the training set;
- introduces physical awareness in the algorithm avoiding the identification of spurious correlation raised in the real data;
- helps to reveal some additional patterns in the data;
- checks if the model follows the expected physical behaviour.

1.4. Modelling approaches

After a small comparison between the two schools of thought most present in scientific know-how applied to system identification, we will identify how the concepts just discussed are implemented in practice. The following section describes the methods and experiences

of different modelling procedures occurring in the landscape of industrial process identification. The choice of the modelling method depends on numerous factors, in particular, the main difference concerns how information derived from prior knowledge and information provided by data is incorporated. For example, we can select a specific kind of identification based on the type of prior knowledge available or how much informative are the data measured. Thus, just by considering this aspect, the choice of identification procedure can take opposite directions and it suggests that it is not appropriate to choose a type of identification method a priori. It is more plausible that using a model that is able to incorporate physical knowledge and extract as much information from the data is more suitable than a model that specialises in one of the two skills. First of all, we analyse the antipodes solutions that are usually shown in identification theory, namely Black-Box and White-Box identification. A first review of the main identification methods will allow us to present a more articulated identification proposal with its related developments.

1.4.1. White-box identification

White box models are completely constructed from the physical insight and internal workings of a component. It requires complete prior knowledge of the original system because it is a mechanistic representation of the kinetic knowledge, as well as on energy and mass balance. Since it explicitly represents the characters of the process it provides transparency in the predictions. So by looking at the internal model parameter, we can visualize the real dynamics of the components. Generically, the white box model explains simple linear and monotonic models and it provides an extreme approximation of the reality. The estimated system provides a lot of interpretability but it isn't able to predict well the real process because the reality cannot be summed up to a first-order system. Since the representation of the principal dynamics doesn't cover the several high order phenomenon, the white box approach lacks accuracy. Furthermore, the system is placed in a noisy environment that affects the dynamics of the real components. The necessity to evaluate high order dynamics, bring the science to interrogate an approach that builds the identification procedure only on the measured data.

1.4.2. Black-Box identification

Black box models only represent the behaviour of a component and are usually learned from data assisted by experience concerning the modelling procedure. They, barely, represent the relationship between observed variables without investigating the physic interpretation of those. Thus, without being able to fully understand their inner workings it is almost impossible to analyze and interpret their predictions. It often needs access to the

original data-set in which the Black-Box model was trained in order to explain its predictions. However, the black-box approach is widely used in industrial process identification because of the arrival of data-driven techniques that make the identification quicker and smarter. The main advantage of the aforementioned approach is that it doesn't require to set a prior the complexity of the problem so it is disposed to identify any type of system with any accuracy level. By relying on sophisticated machine-learning classification models trained on massive data-sets thanks to scalable, high-performance infrastructures we risk overfitting the data trying to identify the disturbances and create a model sensitive to individual correlations. So, even in the data-driven field, there are trade-offs that have to be analysed in order to make the identified system suitable for control. One of many is the conflictual nature between exploitation and exploration. The exploration effect arises when the aim of the identifier is only to find the correlation of data and it put all its effort to better fit the data points. Instead, the exploitation effect turns out when the focus of the overall model is to identify just the main logical correlation so that the identified one is the most general.
When we acquired a minimum amount of knowledge via exploitation, we can progressively enforce the model to precisely find the patterns in the data sets. The outcome of the identification procedure guarantees the existence of a much more accurate model because these models are unconstrained by the physic on the parameters like in a mechanistic model.

1.4.3. Grey-Box identification

From the previous itemization, we conclude that the two main identification approaches have different limitations and the more efficient technique is a compromise between the two categories. Therefore, we introduce the Grey-Box modelling approach that is an ensemble of Black and White-Box models. Thus, the section focus on finding an approach that is both accurate and interpretable in such a way as to achieve the fair trade-off between the aforementioned qualities. It is a cumbersome task because the higher is the complexity, the higher is the accuracy and the lower is the complexity, the higher is the interpretability.

First of all, we can define different branches of grey box identification that differs in term of the way prior knowledge is included in the model. The first category is called constrained Black Box identification where it involved the least amount of knowledge of the real system because it barely constraints the specific parameters in order to not generate inconsistent estimation with respect to their physical realization. So the mechanistic contribution is limited to avoid the presence of inconsistency created by noise or errors in the measurements. It is a straightforward transformation of a continuous-time model into a discrete-time model where the limitations are imposed as constraints in the static gain

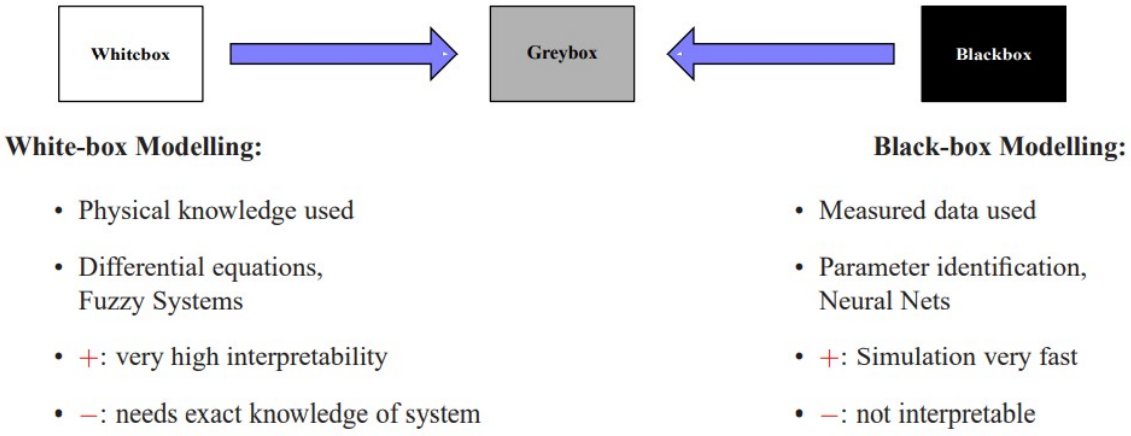

Figure 1.1: Comparison between white-box and grey-box

and in the time constants. When the process exhibits significant nonlinear behaviour, linear black box models give a poor correspondence between the process and the model behaviour. Then, the semi-physical modelling procedure employs the physical insight to create a new variable started from the old inputs and outputs variable. The new elements are used as regressors to create a linear black-box model. It is convenient to use this method only when the first principle modelling is not feasible because if we own more insight into the physics that govern the real system we can take advantage of it. For cases where a fundamental insight into the mechanisms that underlie the behaviour of a process exists, relevant balance equations can be formulated as a set of first-order equations. Even if just a few parameters are well known because, for example, they are common industrial components, the use of a small amount of prior knowledge can lead to a more interpretable model structure. Furthermore, the computational effort taken part by the identification procedure is less because it is limited to estimating only the unknown parameters. In this way, we rely less on data and the informativeness of the data contributes only when the physical knowledge is uncertain[16]. From a practical point of view, grey-box modelling is a very convenient way to model nonlinear processes, since, the model structure can be derived from the first principles of mass and energy balances and the nonlinear characteristics of the process can be modelled as an empirical additive component. Model identification through grey-box modelling is a very effective method because it combines the strengths of a white-box model and a black-box model. On the one hand, the grey-box model possesses qualities inherited from a white-box model, e.g. the need for a lower level/quantity of data. On the other hand, it inherits the potential of a data-driven model, i.e. the ability to counteract physical uncertainties. Certainly, the flexibility of a grey box model is less than a model physically based because the latter

has greater generalisation properties. But this peculiarity does not discredit the use of grey box models as the applicability is superior, just think of the scenario in which the identified model performs monitoring or is used to implement online control strategies. In these cases it is essential to know the physics that governs the real system in order to balance the control structure according to the responses it provides.

1.4.4. Hybrid model identification

Consequently, dealing with grey box modelling is not enough to develop a precise model because it requires selecting in advance the complexity of the model based on our prior knowledge and on the physic that describes the system. Furthermore, the simplification of complex and non-linear phenomena entail the neglect of the high-level dynamics that in turn, neither a physically accurate model is able to reproduce.Therefore, it is recommended to update the model under analysis with a component that uses statistical modelling techniques that are not based on real parameters. Precisely, a physical model is incapable to model either the errors incorporated in the data (data uncertainties) or the structural deficiency that are introduced by the approximation of the physic. By the way, explaining physic phenomena by equations is already an approximation because it just models the main dynamics of the system and there will always be some hidden/high order dynamics that we neglect for the sake of comprehension of physic itself. Moreover, the addition of a data-driven model is contemplated to bring specialization property to the overall system. Generally, we can add the machine learning model in two ways: the first one is the one shown in the figure 1.2, which is the serial structure[11].

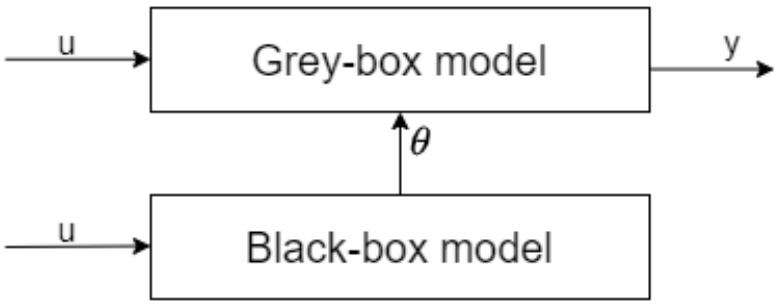

Figure 1.2: Serial structure of a grey-box model with machine learning model.

The serial structure is beneficial when the complete description of the entire process is available. In this case, it brings an enormous advantage to the pure grey box model because the physical variable is updated by a fully physical unmotivated model[15]. The serial structure can take on two architectures. The first approach consists of the initial use

of a black-box model whose task is to estimate those parameters that are not present in the mathematical model (white-box model). Subsequently, the response of the black-box model is injected into a phenomenological model that describes the target dynamically. A second variant of the serial structure contemplates the opposite sequence where the mechanistic model precedes the data-driven model. In this scenario, the parameter that cannot be physically modelled is a function of some component of the first-principles model. The major disadvantage of such a structure is the progressive loss of information of the uncertain components of the system. Therefore the overall model will be less accurate and less interpretable because the patterns that a black box model returns have a high priority, obscure and corrupt the physical dependencies in the data. An alternative is a parallel approach (Fig. 1.3) where the grey-box model and the black-box model are fed by the same input data-set. In this case, the data-based model has the task of modelling only the residual dynamics that a physically-motivated model has not been able to capture because they are of a higher order or not covered by the mathematical equations. [13]

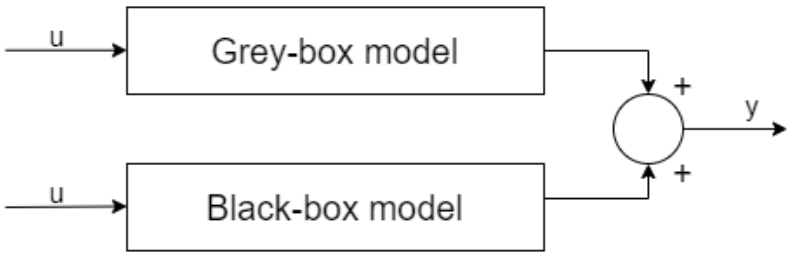

Figure 1.3: Parallel structure of a grey-box model with machine learning model.

Therefore, the output of the hybrid model is the sum of two separate model outputs.

$$y = y_{grey} + y_{ML}$$

In other words, machine learning is used to determine the unstructured uncertain dynamics, that are not represented with their adequate complexity in the first principle equations, inside a grey box model. The black box model has to figure out the residual between the real dynamics and the first principles one so that the behaviour of the overall model fits the reality unless of phenomena caused by noise alone. To mitigate the disadvantages of grey box modelling and pure machine learning modelling, a new approach is considered namely hybrid modelling with limited authority. In other words, the initialization of the parameters is made by an interpretable model as a grey box model in which the physics of the system is written through partial differential equations like mass balance and energy balance equations. Progressively, the unmodelled residual dynamics train a black-box model that tight the gap between the estimated parameters (derived

by a priori knowledge of the system that in turn are been figured out by the grey box model) and the real parameters that have a more complex nature with respect to their physical representation. The choice of black-box structure is limited to linear frameworks since the mismatch between the grey box model and the real system is generally small: ARX models. In this dissertation work the choice of black-box structure is limited to linear frameworks since the mismatch between the grey box model and the real system is small: ARX models. But such an architecture can be extended to empirical models of any nature with the advantages and disadvantages that a complex structure possesses. ARX stands for Auto-regressive with Exogenous Variables, where the exogenous variable is the input term. The ARX model structure is given by the following equation:

$$A(q)y(t) = B(q)u(t - n_k) + e(t)$$

where the predictor depends on the previous output, the previous input and a white noise disturbance. The identification method for the ARX model is the least-squares method, which is a special case of the prediction error method. The least-squares method is the most efficient polynomial estimation method because this method solves linear regression equations in analytic form [14]. Moreover, the solution is unique because the cost function to minimize is always convex regardless of the data-set. It has only a global minimum so it is insensitive to the initial condition.

In addition, in order to not sacrifice the interpretability of the parameters is paramount that the machine learning algorithm participates only when its data are coherent with the physic of the system and does not introduce spurious correlations (it ignores causality). This peculiarity explains why we introduce an authority factor that weightless the contribution of the non-physical model because we trust more to the first principle equation that describes for the most part the model of interest.

1.5. Related work

The following understanding aims to relate the research of this thesis to the overall scenario by considering approaches similar to the described methodology. Given the polyvalence of the subject, the identification of systems by means of a hybrid model covers many sectors, from chemicals to telecommunications, from energy to the environment. It made possible to test an identification methodology in case studies which are completely different from each other and which present specific problems due to both the nature of the system and the availability of data that can be acquired from the system.

Subsequently, I will analyse the previous work trying to identify the main limitations of the methodology used. This will help us to understand if clearly the new methodology we propose is effective in the overall scenario. The papers I have taken into consideration cover a period of about twenty years, which confirms that ideas persist in the community. Taking such deep-rooted ideas into account could be relevant in terms of scientific know-how.

The first scientific paper analysed was in 2002, in which **Qiang Xiong and Arthur Jutan** [11] used a grey-box model in parallel with a neural network to model and control a chemical process. The dynamics of a chemical process is extremely non-linear and has certain characteristics that are difficult to model even using a good non-linear model.

The main reason for this is the strong specialisation of a model to one class of non-linearity, so using one undermines the identification of non-linearities belonging to another classes. Therefore, not knowing what kind of non linearity is predominant of the system they combined a third order linear model with a neural network. The main considerations are of a methodological nature, i.e. they confirm the advantages of using a model based on physics and recommend that it should be combined with an empirical model since in their application the writing of a physically rigorous model is impossible for the reasons explained above.

The responsibility of the neural network is to identify the non-linear effects of the reaction heat on the reaction temperature. An empirical approach requires a large amount of data which fortunately, in the case study they analysed, was easy to find. But contemplating a scenario in which a scientific experiment could take days or years to collect data, prioritising a neural network would be a poor choice. Instead, as Qiang Xiong, Arthur Jutan pointed out, prioritising a grey-box model led to an acceptable simulation of the real process and allowed model errors to be compensated for when the operating point moved outside the linearization region of the approximate model.

The modelling of a phenomenon occurring on a longer time scale has been analysed by **Francesco Massa Gray and Michael Schmidt** [4] who present a hybrid approach for the identification of building heat transfer.

The physics-based component of the model consists of a highly simplified grey-box model where many phenomena have been ignored. The grey-box identification is limited to the estimation of 8 parameters between heat capacity and term resistance as the real system has been represented by means of electrical analogy. The empirical component, on the other hand, is represented by a Gaussian process which has the task of estimating the error between the output of the grey box model and the real model.

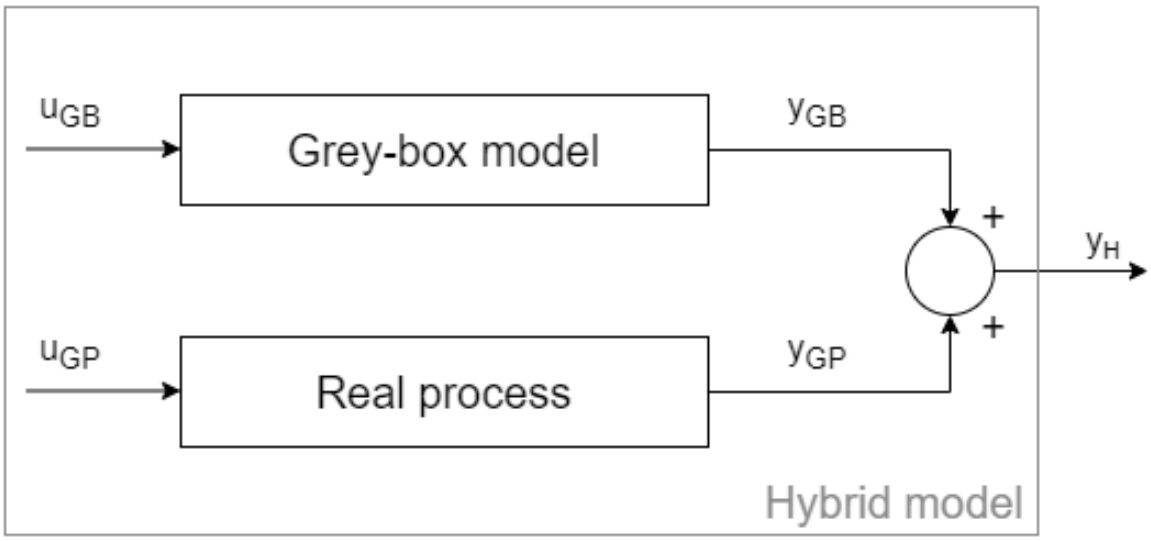

Figure 1.4: Diagram of the hybrid model

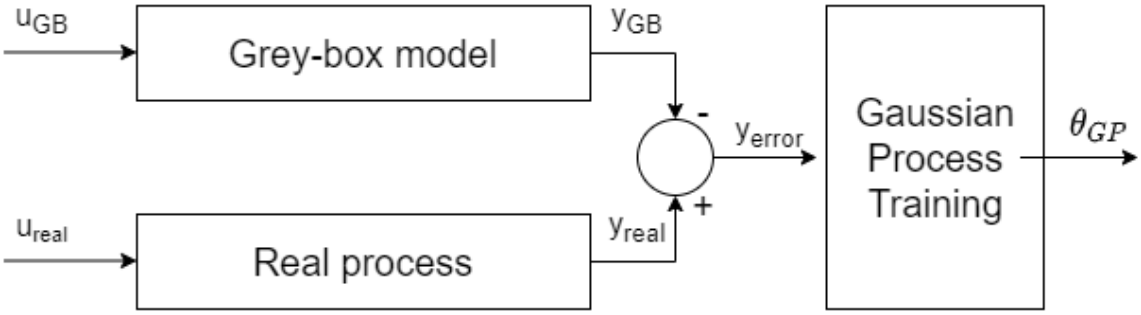

Figure 1.5: Training procedure of the hybrid model's GP.

Francesco Massa Gray and Michael Schmidt recognise the potential of training the two structures separately and not simultaneously as is usually done because otherwise there is a loss of physical interpretability since the stochastic information brought by the empirical model obscures the deterministic information brought by the model based on first principles.

The slowness of the phenomenon under investigation forced the collection of a small data-set and it allowed to evaluate how a hybrid model can generate better predictions than a model not based on physics. Indeed, in the face of small data sets, the hybrid model outperformed the Gaussian process by far, suggesting that it has less sensitivity to untrained data. In short, the presence of a grey-box model acts as a backbone to a hybrid model.

In the area of building thermal modelling, a new paper was published in 2021 in which **Matthew J. Ellis** [3] compares two approaches of hybrid identification. The first one concern the estimation of the main dynamics through a grey-box model and then obtain an empirical model that predicts disturbances through a neural network. The second approach is to simultaneously estimate the high-order and low-order dynamics.

The first approach presents a grey-box model represented by two thermal masses. It proved to be a valid simulator of the physical model as it is able to simulate the thermal zone avoiding a structural error between the model and the real plant. The stochastic error is modelled by a neural network as in a previous example. The use of a black box model of this nature slightly complicated the estimation of the parameters as the parametric identification algorithm is not convex. It was, therefore, necessary to train the neural network with ten different initial values and then verify that they converged to the same value.

In a predictive control scenario, the solution just described is preferable to the adoption of a purely non-linear model as it could be computationally disadvantageous. Instead, the use of a hybrid model allows the controller to be synthesised on the linear part of the model only. It would be pointless to develop a very accurate and responsive controller for this type of application as the dynamics belonging to the plant are slow. The second approach is treated by Matthew J. Ellis in the same way (same neural network).

The final conclusions are approximate because they define one approach better than the other only by analysing the coefficient of variation of the root mean squared prediction error (CVRMSE) without defining the purpose of the model. Because as we have also pointed out above, an identified model can also be used for different scenarios of the pure simulation. He concludes by defining the model generated by the second approach as the best in terms of performance. However, it is deducible that the simultaneous estimation of a grey and a black model guarantees less interpretability to the system because physical knowledge and empirical relations are mixed together. Likewise it limits the modelling flexibility in terms of variable authority between the two methods.

An analysis of the methodology at a more conceptual level is provided by **Z.F. Wu, Jin Li, M.Y. Cai** [19], who discussed the nature of an artificial neural network (ANN). According to their point of view, ANN can be considered a white-box model or a black-box model depending on how much physical knowledge is implemented. They take as a case study a very simple example of a mass positioned in a plane where a force is applied. The system can essentially be described in two ways: - We can describe it by a physical law (Newton's second law). - we can define the relationship between force and acceleration

through a parametric function

The first model will be called a white-box model because the parameters have a physical meaning, whereas the second model will be called a black-box model. However, if we extend the black box model to a model consisting of several equations, where each equation represents a certain configuration, we can see that some parameters are related to physics, because when the experiment evolves, the change of a physical parameter (e.g. mass) corresponds to the change of a coefficient. For this reason, we can call this model a grey-box model because some coefficients have a physical correspondence and some do not.

The conclusion the authors suggest is that defining a system as a black box model means not understanding what the system is about. Because if we investigate the real system, we will surely be able to give it, in part, a physical meaning. Definitely, not all real systems can be interpreted in the same way because some physical representations need to be described by more complex models, but even minimal knowledge of complex processes can improve a model based on data alone. The introduction of a physical component into the model can only benefit and make the next model more intelligent.

The final conclusion to the question of whether an artificial neural network is to be considered a grey-box model or a white-box model derives from the weight of the physical knowledge contributed. If the physical principles are not dominant within the model then the model is defined as grey-box.

At the 51st IEEE Conference on Decision and Control, **Christian Paraiso Salah El-Dine, Seyed Mahdi Hashemi and Herbert Werner** [1] compare knowledge-based models with data-driven models in industrial controls scenario. Their research focuses on black-box and grey-box identification techniques for linear parameter-varying (LPV) systems. A 3 DOF robotic manipulator was chosen as a case study. Subsequently, the manipulator was identified through both a grey box model and a black box model, then, a controller was synthesised for each model. Comparison of the results shows that controllers based on a grey-box model outperformed data-based models in all scenarios except in the independent joint PD controller because it does not require a physical model. However, the implementation of a data-driven LPV-IO input/output controller, combined with a technique that reduces the complexity, was able to match the performance that a grey box model achieved.

The last result obtained, proposes a data-driven methodology able to match a grey-box methodology. The adoption of a black box approach is useful when obtaining a physical model is complicated and expensive, or when the phenomenon to be identified

is so complex that the development of an adequate model requires the development of a too detailed grey box model. This would not allow the synthesis of a controller.

In this instance, the efficiency of a data-based model is preferable even if it is not physically motivated. It should be noted that the presence of physical awareness within the LPV-IO model was not completely ignored as the choice of an acceptable model grade was directed by physical insight otherwise it could not have been assigned a priori.

Support for the use of physically interpretable models was also provided by **Jan-Philipp Roche, Jens Friebe and Oliver Niggemann** [6], who implemented a grey-box learning methodology in an EMC setting. The EMC field has no complex finite-element structures capable of making future predictions, particularly with regard to radiation emission. Therefore, the implementation of a grey-box model is appropriate because it combines a model that can be described by closed mathematical equations (equations that describe electrical components mathematically and physically) and a black-box model that describes phenomena that are unknown or difficult to represent in the form of mathematical equations. They define a grey-box approach as a combination of the advantages of a bottom-up method (white-box) shown in Fig 1.6 and a top-down model (black-box) shown in Fig 1.7. They consider two grey-box structures, both of which insert physically describable parts and data-based parts in a nested manner. The first solution is to consider the external structure as a white-box model and the unknown components as black-box models in order to create a versatile and adaptable structure.

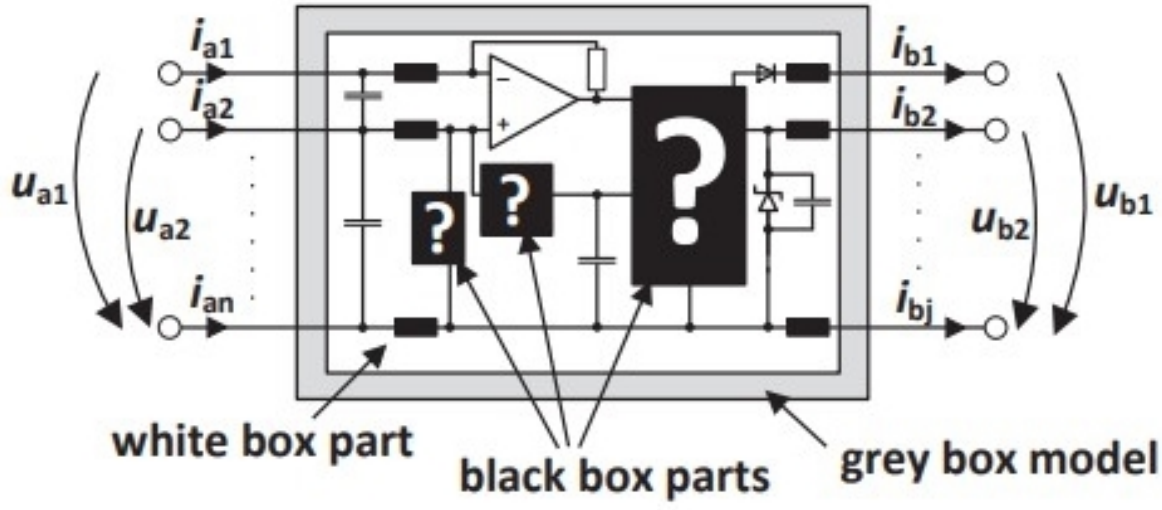

Figure 1.6: Structural white box framework of grey box model

The second approach is defined as dynamic knowledge-based neural networks where the relationships between the physical components are established by non-parametric func-

tions.

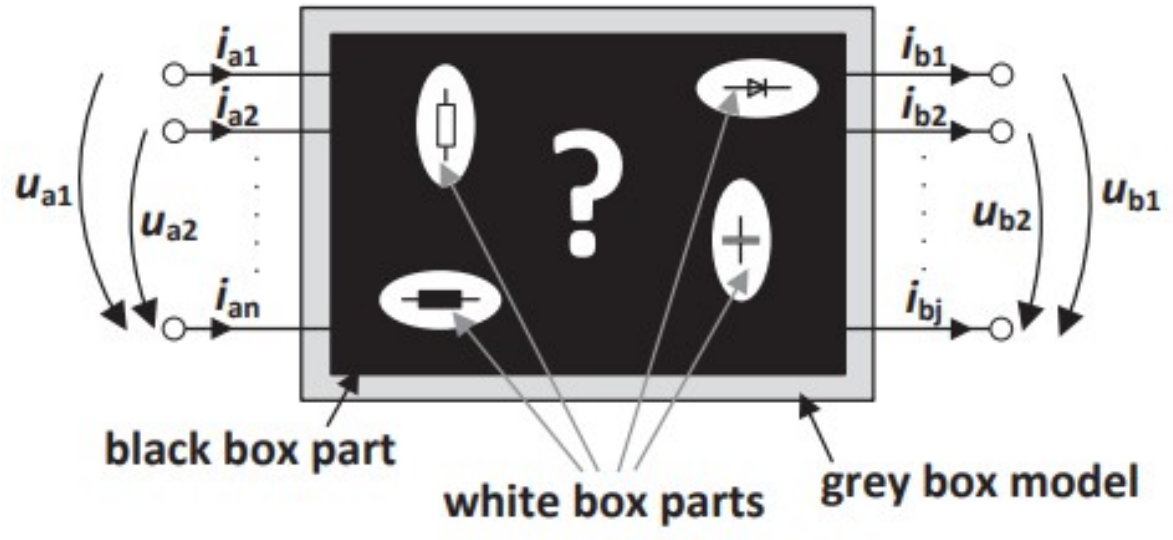

Figure 1.7: Structural black box framework of grey box mode

A nested structure, such as those, models the black box parts and the white box parts simultaneously: a structure is integrated into another and vice versa. The direct integration of heterogeneous models leads to a dispersion of the physical knowledge that one approach (white box) introduces. Furthermore, the identification of the nature of a signal becomes difficult because the white-box model is contaminated by structures that cannot be physically interpreted. Therefore, separation of the identification process is recommended otherwise the advantage of introducing scientific awareness is nullified.

In order to exhibit a complete review of the literature, it is essential to analyse scientific research that does not adhere to the motivations of the thesis to also highlight the limitations that a hybrid approach might present. Especially the introduction of scientific knowledge into the model is linked to the depth of human understanding of the phenomena. Humans are generally more prone to introducing errors within the model so adopting a model that has not interacted with humans during the training phase could be more accurate.

An example that **Christy Green and Srinivas Garimella** [2] analyse is water heaters heat transfer, which are commonly modelled using grey-box architectures. They show that using a pure data-driven model such as ARX is able to outperform a grey-box architecture as it is able to infer unknown dynamics from the grey-box model. The study was also carried out using a more articulated black-box model consisting of two XGA layers (a gradient boosting variant) where the control logic of the heater pump was deduced from the model independently of knowledge of them (Fig. 1.8). The disadvantage of the grey

box model in this scenario is its inability to adapt to water heater types other than those for which it was designed.

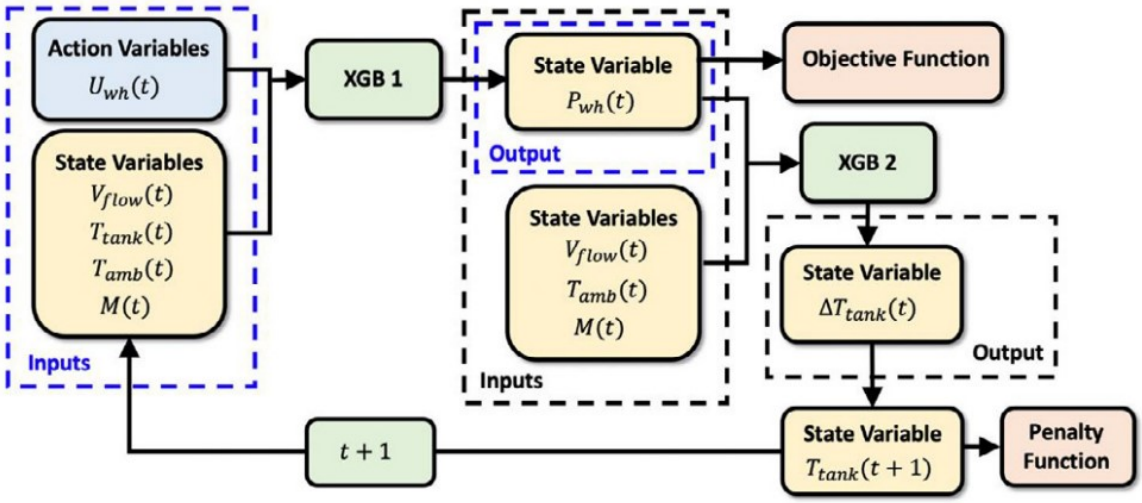

Figure 1.8: Water heater black-box model flow diagram

For example, to calculate the power drawn by the water heater, it is necessary to know the COP performance coefficient, which is different for each water heater. A data-driven structure does not need to know the COP value but just a data set of input and output of the variables to be controlled is required. Since XGA is a recursive structure, it consists of two layers that interact bijectively with each other, the prediction is more accurate because the hypothetical COP value is iteratively updated according to the type of scenario encountered.

The provision of greater accuracy is balanced by greater energy consumption by the black-box architecture, which must be constantly updated in order to capture the transient dynamics of the system.

One area requiring high-speed software is the field of telecommunications where today's development environment needs to be updated constantly. **I. Skuliber, D. Huljenić and S. Dešić** [5] present an update of the development process by introducing new modelling elements into the process. To not create parametric interference between the current model and the newly added model, both systems will be modelled and compared.

For modelling legacy system's characteristics the use of a white box model created through reverse engineering is not effective as it has too many modules so it will be treated as an interconnection of consciously developed black-box models. The new modelling element will be identified through a grey-box approach. The result is the creation of a combined model which is able to predict correctly the experiments carried out on a real architecture.

The implementation of a different kind of model in an extremely fast environment is an opportunity for comparison with the aforementioned literature. The ultimate conclusion is that a complex and parameter-rich system if modelled with a grey-box model could provide an overly complicated and slow model. Therefore, in domains where speed of response is crucial, it is not appropriate to include a physically motivated model as it is too lazy. It is smart to drive a black-box model with expertise where the training data-set is the most smart and rich of relevant information possible.

A comprehensive review of hybrid methods was carried out by **Timur Bikmukhametov, Johannes Jäschke** [17] where they list multiple approaches in which the mechanistic nature of a grey-box model can improve the performance of a machine learning model. In order to be as explanatory as possible, a process of producing oil and gas from a well located in an oil production system was chosen. The use of a model based on physical components is impossible to contemplate as the flow is multi phase and would further require a very articulated model that takes into account losses during extraction and many other phenomena. The implementation of a model based solely on data such as feed-forward neural networks, gradient boosting and Long-Short-Term-Memory (LSTM) neural networks has been considered but the output data it returns is difficult to interpret.

Furthermore, the task of a machine learning model is to reveal how system parameters relate to each other so if we introduce raw measurements directly to a data-driven system it will also try to combine parameters that do not describe the dynamics of the system. If instead, the raw measurements are initially processed by a structure that produces physically compatible data, the machine learning model will be able to generate true and meaningful parametric matches.

The most accepted approach, therefore, is to combine first-principles models with machine learning in order to take advantage of both methods. There are five approaches:

1. Future engineering: the simplest model in which the measurements provided as input to the machine learning model are pre-processed in such a way that they are interpretable and based on process knowledge. (Fig. 1.9)

2. First principle model solutions and feature engineering: is an extension of model 1 in which the machine learning model inputs are generated from the error between a first principle model and the target value.(Fig. 1.10)

3. first principle solution and raw measurament: similar to model 2 where as input to the machine learning model the raw measuraments are inserted directly. (Fig. 1.11)

4. Linear meta-model of models with created features: can be summarised as a weighted

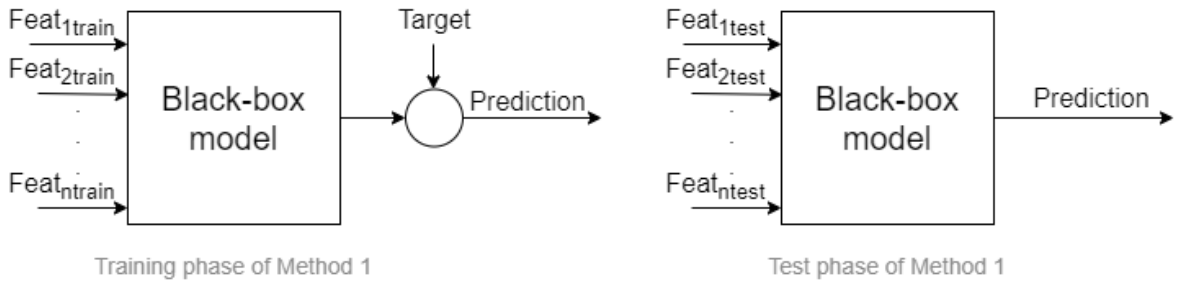

Figure 1.9: Training and test procedures for Method 1 - feature engineering

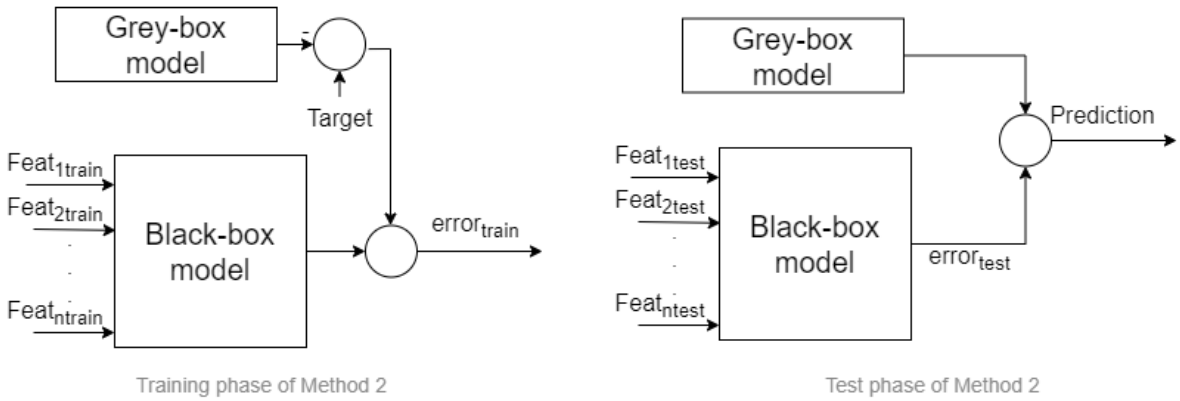

Figure 1.10: Training and test procedures for Method 2 - first principles model solutions and feature engineering

superimposition of machine learning models. Each model represents a particular component of the oil production process. (Fig.1.12)

5. Linear meta-model of selected model with created features and model with raw data: summarizable to a weighted superposition of models of type 1. (Fig. 1.13)

Disclaimer: In order to have comparable results, all models have been tuned by the same optimisation algorithm (Bayesian optimisation). It is important to remember that if we use different optimisation algorithms, the generalization property may vary because one particular algorithm could be more accurate than another in a specific scenario.

The conclusion that all approaches have in common is that a hybrid model is more accurate, interpretable and generalisable than a pure machine learning model. Furthermore, the use of raw measurements in a hybrid model benefits the identification of a data-driven model while maintaining explicability. This statement is only true in case of a hybrid model, in case of black-box models it is impossible to understand if the results have a physical meaning or not because they are too corrupted by noise and uncertainties.

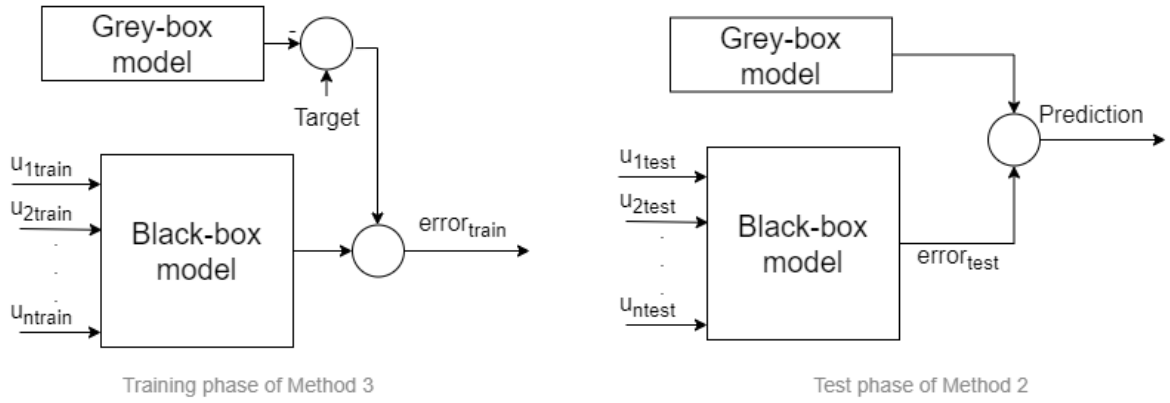

Figure 1.11: Training and test procedures for Method 3 - first principles model solutions and raw measurements.

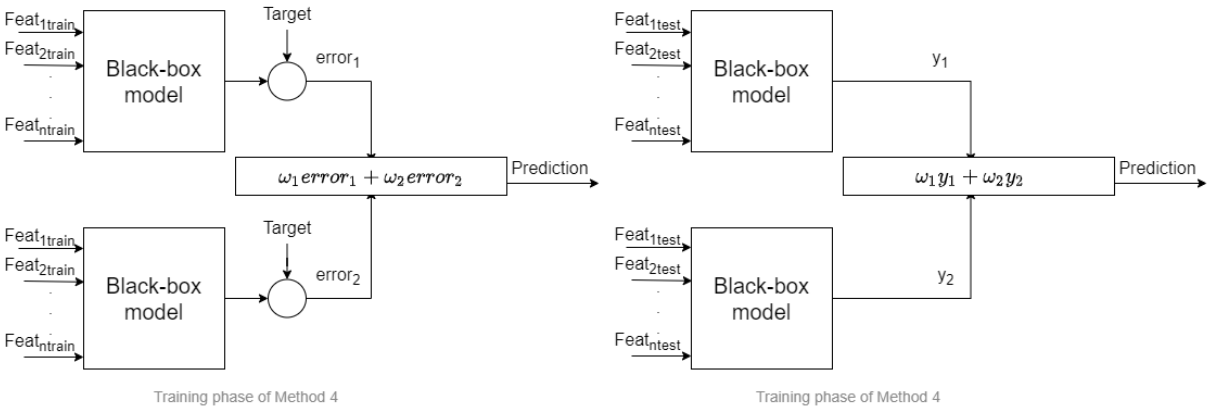

Figure 1.12: Training and test procedures for Method 4 - linear meta-model of models with created features.

A further consideration arises during the a posteriori analysis of the features because they allow us to verify, firstly, whether the feature created is meaningful and mimics the component we wanted to represent and secondly, whether the model based on first principles follows the expected behaviour. Similarly, having a transparent system allows us to compare grey-box models of different natures, establishing which of the selected ones is best able to explain the data and reveal patterns that have not been taken into account.

In terms of performance, the model that stands out is model 5 because has the capacity to include a lot of information about the physical components but simultaneously it is difficult to build because a high level of physical knowledge of the phenomenon is required. Models 2 and 3 are, however, effective in contexts of moderate complexity, they presented some difficulties in estimating parameters that vary irregularly over time. This may be

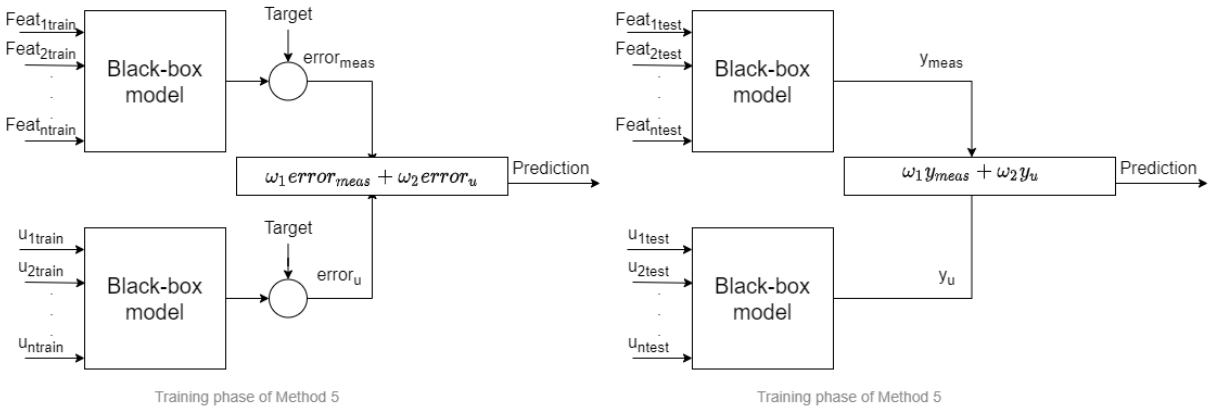

Figure 1.13: Training and test procedures for Method 5 - linear meta-model of the selected model with created features and model with raw data.

due to the fact that the physical modelling of the first-principles model considered too superficially the phenomena that are most difficult to identify. A summary of the model selection and on the consideration that has to be done if we want to integrate a machine learning model into a knowledge-motivated one is show in the figure

The authors remind us that increasing the complexity of a physical model will undoubtedly create a more accurate model but concurrently it will be computationally less efficient.

1.6. Conclusion

The analysis of the literature highlighted the know-how present in the scientific community. Through a heterogeneous choice of papers, I was able to illustrate in a comprehensive way the scenario concerning the identification of systems through a hybrid structure. In addition, the authors mentioned several times the advantages that the incorporation of first principles into a model can bring. In particular, the last paper analysed, presented a comparison of the methodologies used in the field of hybrid system identification. The methodology that comes closest to the one we studied is number 3, which uses a physics-based model to identify the main dynamics of the system and incorporates an empirical model to identify the residual dynamics.

One aspect that is little taken into account is what we call "variable authority". The scientific community deeply investigates the role of the grey box model and the black box model separately. Most papers focus on the development and analysis of a data-based model assuming a first-principles fixed-order model. In other words, an approach in which the degree of complexity of the grey-box model is assessed according to its purpose is less well taken into account. The implementation of multiple grey-box models of different

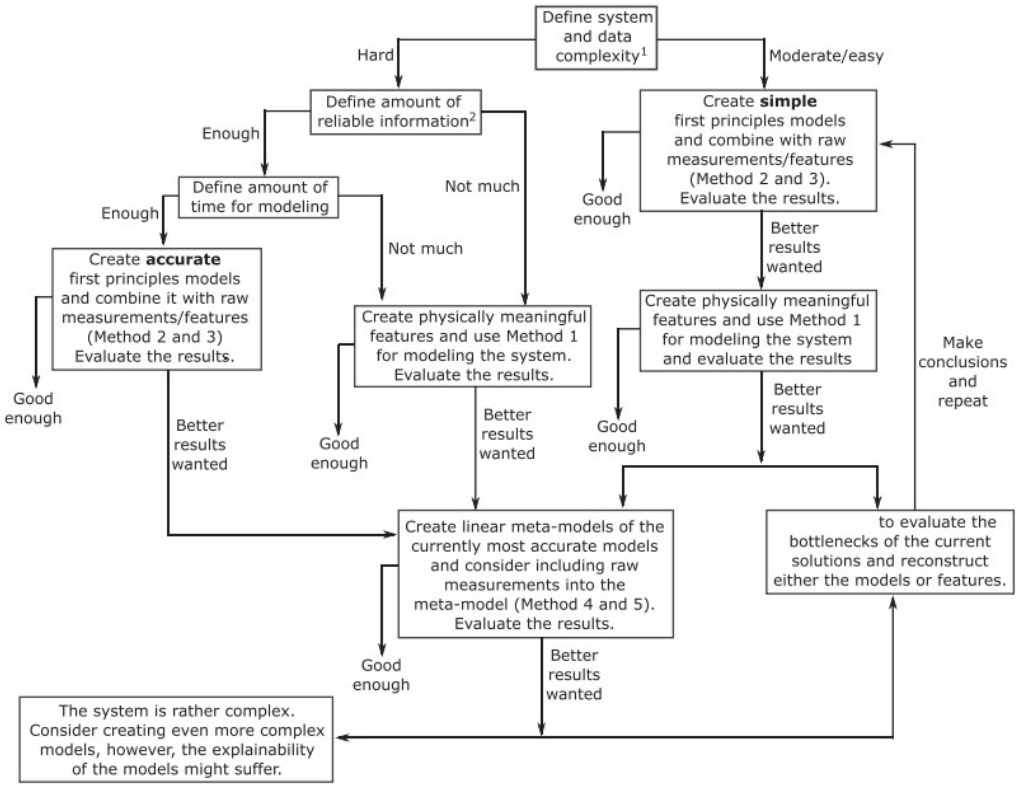

Figure 1.14: Summary of the method selection for process system modeling

complexity, implicitly, imposes a different authority on the data-based model.

To summarise, the accuracy of the grey-box model, in a hybrid identification context, depends on the physic and on the purpose for which the identified model was created. Thus, the grey-box model will have the task of identifying the dynamics of the real system according to its potential. If it is extremely simple, the model based on first principles does not have the tools to model higher-order dynamics or uncertainties in the model/parameters. In this case, the intervention of a data-based model will be greater because the residual still contains significant dynamics of the real process. On the other hand, the implementation of an extraordinarily articulated grey-box model will preclude a high influence of the empirical model by giving greater interpretability and generalisability at the expense of accuracy and lower computational demand.

The early imposition of authority limits the versatility that a hybrid model is able to render.

2 | The proposed approach

Chapter 1 introduces the know-how on system identification. In particular, we focused on discussing why the interpretability of a model can bring several advantages to the identification and the practical necessities of using it. Successively, we described how the data-driven models are not well suited to provide explanations on the physic phenomenons due to their stochastic nature. Then, we examined the main identification approaches in the sphere of the mechanistic models and of the empirical models. Finally, we proposed our identification technique in which the stochastic and the deterministic field are bounded together with variable authority. Now, we define thoroughly the identification method including the selection of the authority based on the physic and the purpose of the grey box model.

2.1. Grey-box model complexity

Advanced modelling tools have enabled the development of detailed complex mathematical models that yield reasonable and accurate predictions of the behaviour of any type of (process) system. The intrinsic complexity of processes yields models that generally require a considerable computational effort to solve them. One of the origins of complexity is the modelling of the distributed nature of physical processes, which results in models containing ordinary differential equations (ODEs). The model order is defined as the number of first-order differential equations.

Order selection and model type must be chosen by balancing two factors.

- The first factor is imposed by the physics of the system under analysis: the nature of a process is heterogeneous since it can be more or less complex depending on the process itself. Every real system has a minimum order of complexity to be represented. For instance, let us consider the identification of a heat exchange process between a fluid and a surface in a turbulent flow regime. In order to describe the phenomenon accurately, the associated grey-box model must present, in the form of mathematical equations, an expression defining the heat exchange coefficient,

since it is essential for describing the event. In turn, if the model has to perfectly mimic the heat transfer dynamics, it will introduce a description of the heat transfer coefficient, which could be the Dittus-Boelter correlation.

- The second factor is imposed by the purpose of the identified system: indicatively, two main purposes can be distinguished. The model may have simulation purposes or control purposes. A model with a high order contains a high amount of physical insight with respect to a low order one because it is able to extrapolate more detailed physical patterns. This scenario imposes building a complex control and a relatively computationally expensive structure especially if the process model equations are invertible. Therefore, the adaptation of the model structure to the control structure is essential and constraints the amount of prior knowledge that the mechanistic part examines. On the other hand, a very simple model (linear or first-order) can lead to a superior control performance because it has an extrapolative character even if it is fewer parameters explicable, might be difficult to adapt and owns less generalization property. A fortiori, if the aim is to tune a linear controller (PID), the model must be linear, so consideration of higher-order models is out of the question.

The trade-off between these two factors limits the development of a highly detailed model, so reducing the order of mathematical models is a route that must be taken in any circumstances.

Therefore, this necessitates the reduction of a system consisting of a large finite number of equations to a system consisting of a smaller number of equations or to a system containing a smaller number of first-order differential equations. This procedure is generally referred to as model reduction and it is contemplated for almost any automation field until the relevant dynamics of the original process are evaluated.

The motivation to perform a model approximation is to improve the computational efficiency while keeping desirable model properties intact. However, a massive order reduction could not allow the extraction of relevant features for the specific purpose for which the model is meant. Therefore, before diving into the identification procedure is appropriate to choose a plausible order of the model based on the physic and on the purpose. In this respect have some prior knowledge of the real process is useful because it can facilitate the analysis of the identification itself.

At this juncture the structure of the grey box model is untouchable since it reflects the perfect balance between the two factors listed above (physical and purpose). So from now on the authority of the physically motivated model is fixed. However, if we use the model to identify the real system we notice that it is not as accurate as a more detailed model

even though it incorporates all the advantages that a physical model can bring: from generalisability to interpretability to reliability. The cause of a lower accuracy can be due to factors concerning the construction of the model itself such as choice of order, imposition of a wrong physical structure or excessive approximation of the process, or factors related to the informativeness of the injected data. These two reasons can generate a model that is not totally faithful or even unfaithful to reality, as its potential is constrained by the context in which it will operate.

Based on the error between the simplified model and the real system, the introduction of a non-physically-motivated structure is considered in order to compensate for the shortcomings that a model based on first principles presents.

2.2. Empirical model complexity

In order to overcome the uncertainties of the model, the parallel addition of an empirical model with high interpolation capabilities is considered [10]. Since the extrapolation of the main dynamics was carried out by the grey-box model, we are looking for a method that is able to improve the output of the model without changing its structure. In this way, the influence of the data-based model in the overall model has less authority because it is dedicated only to improving the output signal without introducing more parametric complexity into the structure.

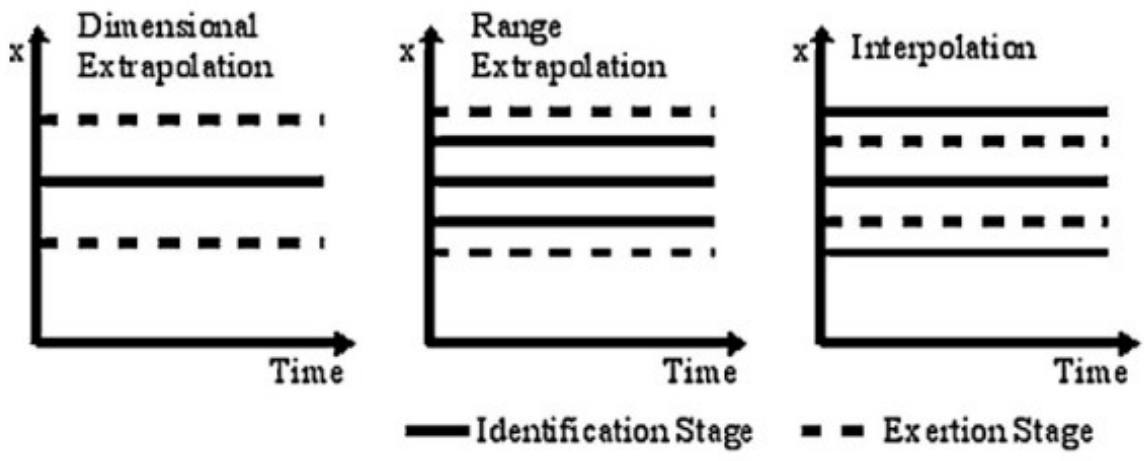

Figure 2.1: Dimensional extrapolation and interpolation

The classification of the non-parametric model structure can be addressed with mathematical and statistical criteria that indicates the most suitable sized for the empirical model. In general, the estimation quality must be balanced against the number of in-

volved parameters and against the number of data that are available for the identification. Furthermore, an analysis of input and output relationship is advised before proceeding because it is a computationally inexpensive tool that allow to understand if exist a dependency between the input and the output. The type and the strength of the relationship carry the choice of the non parametric model structure. For instance, a linear correlation exists when all the points are plotted close to a ipothetic line in the respective parametric plot. The proximity measure of the points designates the strength of the relationship.

Understanding the shape of the parametric plot improves model selection strategies: it gives us a clue to the complexity that the non-parametric model needs. First, the model complexity refers to the capacity of the model in expressing or approximating complicated distribution functions. Its meaning is captured by the notions of model expressive capacity and model effective complexity [18]:

- expressive capacity captures the capacity of the empirical models in approximating complex problems. Informally, the expressive capacity describes the upper bound of the complexity of any model in a family of models;
- effective model complexity reflects the real complexity of the functions in a fixed parameters environment.

A practical example that demonstrates the distinction and relationship between model expressive capacity and effective model complexity is the following. Considering a polynomial function $f(x) = ax^2 + bx + c$, its expressive capacity is quadratic because it is capable of representing at most parabola. When we assign different values to the parameter a, b and c, the corresponding effective complexity changes. In particular, if we choose $a = 0$, $b = 1$ and $c = 1$ the effective complexity becomes linear which is obviously lower than the expressive capacity. To sum up, the expressive capacity can be regarded as the upper bound of the amount of knowledge that a model architecture can hold. The effective model complexity is concerned about, for a specific model, a specific training dataset, how much knowledge it actually holds. In general, model selection and design is based on the trade-off between prediction performance and model complexity. On one hand, making predictions with high accuracy is the essential goal. A model is expected to be able to capture the underlying patterns hidden in the training data and achieve predictions of accuracy as high as possible. To this extent, a model with more parameters and higher complexity is favoured. On the other hand, an overly complex model may be difficult to train and may incur unnecessary resource consumption, such as storage, computation and time cost. To this extent, a simpler model with comparable accuracy is preferred over a more complicated one. In other words, the over-parametrization of an empirical model is

Type of model	Model structure	Parameter estimation
		Static
Linearly	Linear function	Linear regression (Least Squares method)
Nonlinear	Polynomials	Linear regression (Least Squares method)
	Any nonlinear function	Iterative process, Levenberg Marquardt
		Dynamic
Linearly	Transfer functions models(ARX,ARMA,etc)	Linear regression (Least Square method), an iterative procedure
Nonlinear	Neural Networks (sigmoid,wavelet,radial basis networks)	Damped Gauss-Newton backpropagation
	Polynomials (Wiener /Hammerstein model, Volterra model)	Linear regression (Least Square method),

Table 2.1: Black-box model overview

a waste of resources because it doesn't add any advantages.

At this stage, choosing a suitable model structure is a prerequisite before estimating its inner parameters. In this work, we will focus on linear parametric models. A parametric model structure is also known as a black-box model, which defines either a continuous-time system or a discrete-time system. There are a few structures of the model that can be used to represent certain systems. In this study, we will take into account simple linear models with single-input single-output (SISO) like ARX model (Autoregressive with external input) and ARMAX (autoregressive moving average with external input) but for the sake of completeness, I gave an overview of the various approaches in table 2.1.

Firstly, the ARX model structure is one of the simplest parametric structures. The structure of the ARX model can be written in the form of the equation:

$$A(q-1)y(k-n) = q^{-d}B(q^{-1})u(k-n) + e(k)$$

The ARMAX model structure is similar to ARX structure, but with an additional term, which represents the moving average error. ARMAX models are useful when dominating disturbances that have enter early in the process. The structure of the ARMAX model

can be written in the form of the equation:

$$A(q-1)y(k-n) = q^{-d}B(q^{-1})u(k-n) + C(q^{-1})e(k)$$

Whether it's the ARX or ARMAX model $A(q^{-1})$, $B(q^{-1})$ and $C(q^{-1})$ are polynomials to be estimated. For the ARX model, the polynomial $C(q^{-1}) = 1$.

The relevant advantage of AR to ARMA is computational. The optimization procedure in AR model is type 1, which means that the performance index is a quadratic function of the parameter vector. Then the optimization problem has one and only one solution and it can be found explicitly and in one shot. In ARMA process the performance index is a non-quadratic and non-convex function so we need an iterative procedure that iteratively find the local minimum and try to reach the unique global minimum. The procedure is more laborious and doesn't guarantee the attainment of the global minimum.

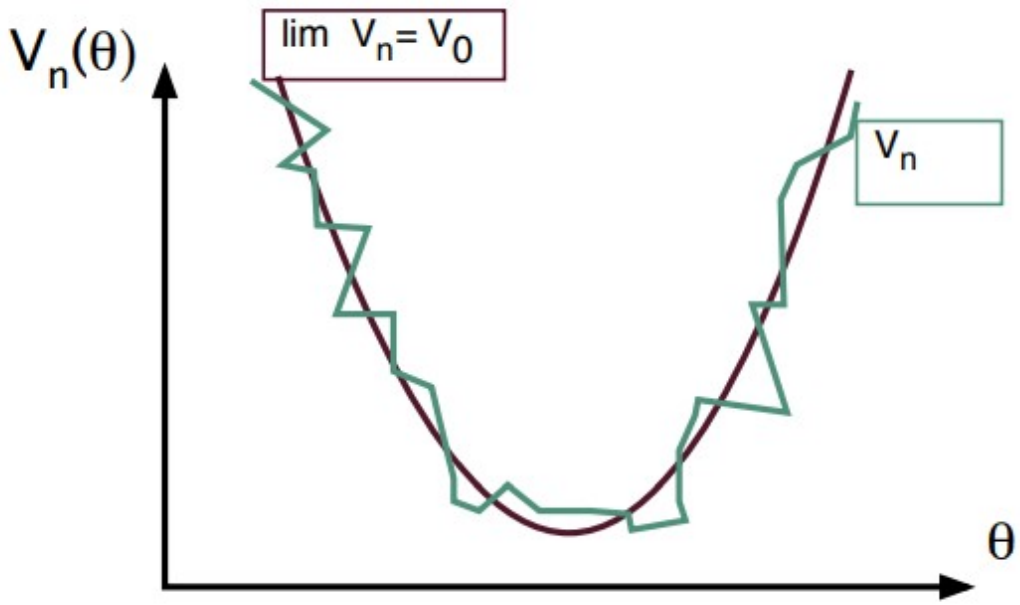

Figure 2.2: Optimization of type 1

The fitting of AR models is essentially a subject of multiple decision procedures rather than that, of hypothesis testing. The procedure takes a form of a sequence of tests of the models starting at the highest order and successively down to the lowest order. To apply the procedure to a real problem one has to specify the level of significance of the test for each order of the model. Although the procedure is designed to satisfy certain clearly defined conditions of optimality, the essential difficulty of the problem of order determination remains as the difficulty in choosing the levels of significance. Also the

loss function of the decision procedure is defined by the probability of making incorrect decisions and thus the procedure is not free from the logical contradiction that in practical applications the order of the true structure will always be infinite. This difficulty can only be avoided by reformulating the problem explicitly as a problem of approximation of the true structure by the model. The introduction of the prediction error allows establishing the set of parameters with the highest levels of accuracy of the predictions computed from the observations.

$$\epsilon(k, \theta) = y(k) - \hat{y}(k \mid \theta)$$

Where, ϵ is the prediction error resulting from y(k) the observed output, $\hat{y}(k \mid \theta)$ the predicted output and θ is the vector of unknown parameters. The Least square algorithm is a common method used in linear system identification, that minimize the prediction error criterion aforementioned. It can be written as:

$$\hat{\theta} = \min_{\theta} \; V_n(\theta, Z^n)$$

Where

$$V_n(\theta, Z^n) = \frac{1}{N} \sum_{k=1}^{N} (y(k) - \hat{y}(k \mid \theta))^2$$

The mere analysis of the correlation between input and output is limiting in determining the degree of the empirical model as it only shows us the dominant trend present in the data. A static feature that does not justify the implementation of a non-parametric model. The first tool that allows us to assign the right order of complexity to the model is the adoption of evaluation criteria. They choose the best fitting order considering two data sets (training and validation) in order to avoid overfitting phenomena. The use of a validation data set for the choice of order is fundamental, otherwise it would be more advantageous to consider extremely high orders when the predictions on the training data set would be optimal (the higher the order, the higher the fitting, the lower the generalisability, the lower the applicability).

There are several approaches that compensate for the automatic decrease of the loss function. Probably the best-known technique is Akaike's Final Prediction Error (FPE) criterion and his closely related Information Theoretic Criterion (AIC). Both simulate the cross-validation situation, where the model is tested on another data set.

$$FPE = V_n(\theta, Z^n) \left(\frac{1 + \frac{d}{n}}{1 - \frac{d}{n}} \right)$$

Where $V_n(\theta, Z^n)$ represent the loss function for the studied structure, d is the total number of estimated data and N is the length of data record. The AIC is formed as:

$$AIC \approx log\left[V_n\left(\theta, Z^n\right)\left(1+\frac{2n}{N}\right)\right]$$

According to Akaike's theory, in a collection of different models, choose the one with the smallest FPE (or AIC).

2.3. Procedure

In this section, I accurately describe the proposed methodology step by step. The procedure begins by identifying a real physical process or a highly accurate simulation (which cannot be analysed because it is too complex) to be modelled and stating the purpose for which the identification is required. Next, I choose an input signal with high extrapolative capabilities so that the excited system returns informative data. Then I define a figure of merit that defines the quality of the model in order to compare them. Through non-linear fitting techniques, I identify the grey-box model. I analyse the residual through the study of the I-O correlation and choose a quasi-linear model that fits well the residual dynamics present in the error system. Finally, I evaluate the error by means of an empirical model and make my considerations on the final result, asking myself whether the use of a data-based model is indispensable.

2.3.1. Choice of the real process

As already announced, the proposed approach goes beyond the choice of the actual application. The choice of the process only influences the minimum complexity imposed by the underlying physics of the process but is not relevant for an effective approach proposal. The approach that I will present will take as a case study a thermodynamic application and an electrical one because these are processes that I have studied during the last years as they are related to the degree course. The introduction of case studies is fundamental in this type of approach because, as we shall see, it is difficult to propose a methodology that is valid for all scientific fields since each system requires different considerations and different choices to be made during the course.

The generalisation of the approach is trivial if one has a thorough knowledge of the field in which it is to be implemented since the interpretability of the data is only possible if the user is able to recognise the physics that governs the given process. In fact, for the sake of clarity, we have decided to collect data from a very complex simulated system

in such a way that the data is akin to physics. On the contrary, if we had considered a real system, the data collected would not be fully interpretable because the process is immersed in an environment that has multiple sources of error (from instrumental error to environmental disturbance etc). In addition, they could introduce spurious correlations that are products of the experiment and not of the physics of the process.

The software we have adopted to produce synthetic data is OpenModelica. It is an open, object-oriented, modular modelling language of an acausal nature. This feature allows one to focus on building the model without worrying about getting a computer executable equation set ready for simulation since this task (called partition generation) is automatically performed by Modelica algorithms.

2.3.2. Choice of the data-sets

Once the simulated model is able to provide synthetic data, the next step is to select a type of input signal that makes the output carry a high amount of physically motivated information. The selection of the signal type is more relevant for the data-driven modelling approach because it is not constrained by physics but relies only on the exported data set. The measurement of the data includes the response to the effect of the input and the sensor noise. Excitation of the system with the appropriate input must be carried out in such a way that the system output is greater than the sensor noise. The quality of the perturbation of the input signal determines the actual change in system response so should be chosen a signal that excites most of the dynamics of the process. The use of a pseudo-random signal such as PRBS is advisable because it fully excites the system around a working region. It can then be combined with classical signals such as steps or ramps to explore different working regions. It is not advisable to use a signal incompatible with the physics of the system because it would excite dynamics that the real system does not present or are not significant and would therefore be a mere request for computing power useless for the software.

Secondly, the acquisition of data provided by different experiments is fundamental to avoid that the model works correctly only on the sample of data used for the training. Therefore, the training data-set is used to train the model and allow it a necessary amount to correctly estimate the parameters but then the model will have to be tested with a different data-set (validation data-set). Generally, the validation set contains samples with known provenance, which allows the operator to assess not only the accuracy of the model but also to make considerations about the dynamics of the response. Based on general guidelines, a step signal is recommended because it provides an interpretable view

of settling time, rise time, presence of overshoot and offset; in short, all qualities related to the temporal behaviour of the model.

However, even following this procedure it is still impossible to tell the performance potential of the predictor on a blind data set because the true distribution of the actual data is not known at the data acquisition stage otherwise it would be data sniping. In real-world applications, the input is normally unknown and one must assume that the performance measured using a blind test set is an unbiased and accurate estimator for the performance of the model on all unknown samples from the same distribution of the training/test data-set. Clearly, without sampling the entire domain, this is unlikely, but it is assumed that with the introduction of a third data-set, the approximation of the system dynamics is well described by the model with a good generalisation property.

2.3.3. Choice of the model complexity

Acquiring as much information as possible is beneficial because it allows us both to gain a better understanding of the actual process and to create an effective physical model. Therefore, it is advisable to plot the dependency of the input on the output in a parametric graph because it gives us an idea of the type of relationship between input and output. If the correlation is linear, a carefully calibrated linear model can mimic the main features of the process. If the correlation is of a higher order we will have to consider more complex models.

Now let us collect the resulting information:

- the physics of the process;
- the input-output correlation study;
- the purpose of identification.

The right compromise between the 3 factors mentioned will define the complexity of the model and the potential accuracy it will be able to achieve. From now on, the structure of the mechanistic model is defined and in turn its authority over the hybrid model is also selected. The chapter will apply the explained concepts by simulating three different scopes of identification in order to have a complete view of the methodology. We will deal with the following scenarios:

1. the hybrid model is intended for the tuning of a controller;
2. the hybrid model is intended for approximate simulation of the real process;

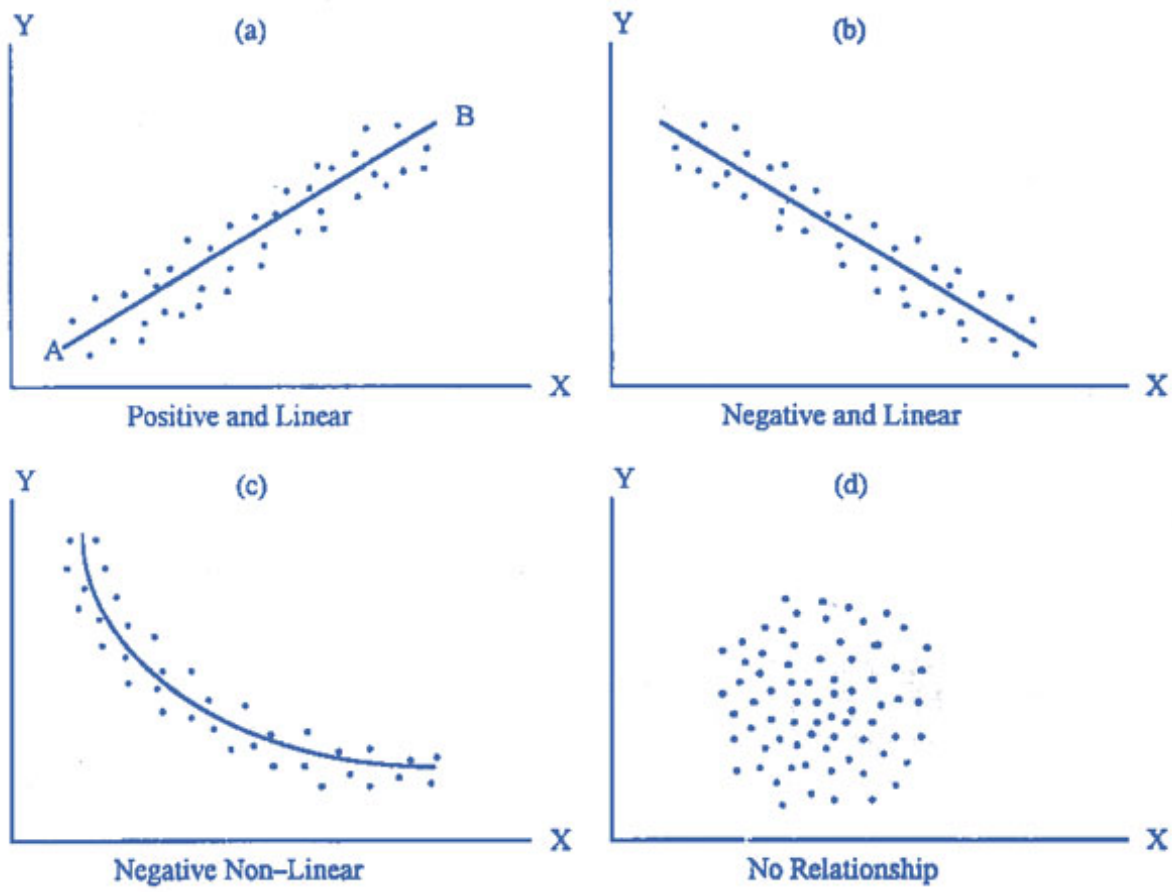

Figure 2.3: Different type of I-O correlation

3. the hybrid model is intended for detailed simulation of the ideal process, also considering non-linear dependencies within the model.

2.3.4. Choice of Grey-box fitting algorithm

Following the procedure, the next step concerns the choice of non-linear fitting techniques adopted for the parametric identification of the grey-box model. However, it is first essential to define a figure of merit that shows the performance of a model and allows comparison between the identified models in order to define the best candidate. Previously I introduced two quality criteria for the choice (FPE, AIC) which will only be used to choose the best empirical model order. To assess the accuracy of a model to the real system, in the form of a fitting metric, I used NRMSE, i.e. the normalised version of RMSE (normalised mean-square error). The RMSE criterion compares two data sets and provides a measure of the error between the two in the form of a residual.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(y_{real}(t) - y_{sim}(t))^2}{T}}$$

$$NRMSE = \frac{RMSE}{\overline{y}_{real}}$$

where y_{real} is the data exported from the real system, y_{sim} is the data provided by the identified system and $\overline{y}_{real}$ is the mean value of y_{real}.

The introduction of a new figure of merit was a choice of convenience as Matlab's *compare* function uses the above fitness value indicator to assess how well the identified model is able to mimic the real system. In particular, the *compare* function returns a percentage fitting value of the NRMSE defined below:

$$fit = 100(1 - NRMSE)$$

Let us return to the subject of this section, namely the description of the fitting methods used to identify the parameters of the grey box model. Depending on the structure of the model, several methods have been evaluated, in particular we can identify three.

The **first method** is typical if the purpose of the identified model is to tune a controller. In this scenario, it is assumed that the model can be defined by a 'gain' K and a single time constant T (first order model), so the identification operation is concerned with finding the values of these two parameters. The simplest thing to do is to apply a step input, with recording of the output trend: the ratio between the steady-state value of the output and the amplitude of the input step determines$K$, while the tangent line to the output curve at its origin makes it possible to determine T.

A fairly simple method is the evaluation of the areas in the recordings as a function of time of the input and output signals of a block: applying a step of amplitude u to the input, the output signal y will have an exponential trend (theoretically $y = u(1 - e^{\frac{-t}{T}})$. Since the integral of the error, for long enough t (at least $t > 5T$) tends to the value uT, T can be obtained simply by dividing the integral of $(u - y)dt$ by u. Note that the error integral is defined as:

$$\int_0^\infty e^{\frac{-t}{T}}\, dt = T$$

This procedure is simple but may be too coarse due to the inherent inaccuracy in the evaluation of the asymptote and tangent. A more precise method (**second method**) is the search for the most approximate curve passing through N points. Let us suppose

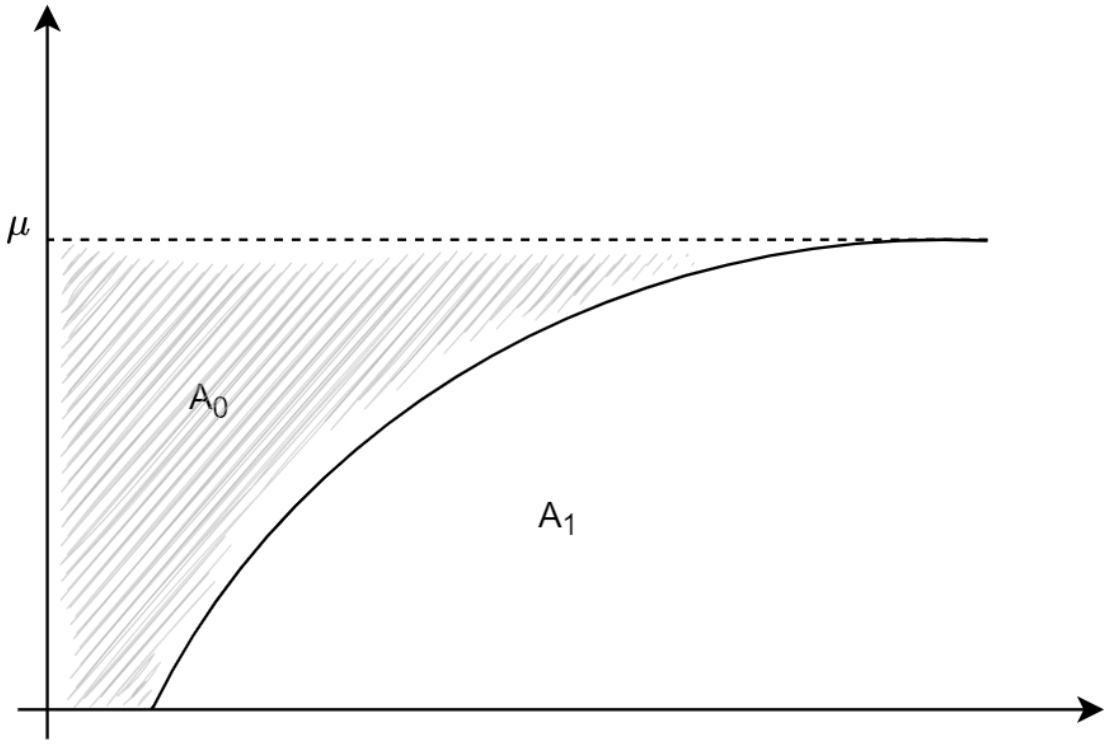

Figure 2.4: Method of the areas

that we have carried out a series of N pairs of measurements of two quantities x and y, which are supposed to be linked through a function $y = f(x; a)$, where a means a set of M unknown parameters, corresponding to as many physical quantities that we want to determine: carrying out a fit means giving an estimate of their "true" value a_0, looking for those values $\hat{a}$ for which the "distance" between the measured y values and those calculated through f from the corresponding values of x is smaller. The estimation of physical quantities is nothing more than an indirect measurement, which will therefore lead to an error at $\Delta a = \hat{a} - a_0$.

The least-squares method is usually used, where the "distance" corresponds to the squared deviation between the function and the measured values.

$$d_i = y_i - f(x_i, a)$$

The quantities are called deviations or residuals: in general they can be positive or negative, so they are not good for defining a distance, which is always positive. We can, however, from the residuals, define a "distance" between the function f and the observed

values of y, by taking the square and constructing the sum:

$$D^2 = \sum_{n=1}^{N} [y_i - f(x_i, a)]$$

This distance, thus defined, depends only on the parameters a. It is evident that, if the function reflects the true link between the variables, then at the "true" parameters the residuals will be equal to the experimental errors, and hence at zero mean and standard deviation y . We will therefore assume that the values $\hat{a}$ at which the summation reaches the minimum are the best possible estimate of the "true" parameters a_0. In general, given any function, the search for the minimum can present a number of problems, due to mathematical reasons (the presence of several minima, for example) or physical reasons (when the value of the parameters corresponding to the minimum has no physical meaning, for example).

An interesting case, however, is when the function f is a polynomial:

$$f = \sum_{k=1}^{M} a_k x^k$$

In this case, deriving with respect to various values of a_k, one always obtains equations that are easily solved numerically. The relation may be constant (of the type $y = a_0$, i.e. there is no dependence on the variable x. Or it can be linear (a line $f = a_0 + a_1 x$ where the value of a_0 represents the intercept of the line with the y-axis, while the value of a_1 represents its slope.

And so on, we can consider more articulated types of correlation (quadratic, cubic, etc.) which, because of Taylor's theorem, should always constitute a better approximation to the unknown experimental law. Unfortunately, this is not the case: due to experimental errors, polynomials of excessively high degree have the tendency to become unstable, and to present characteristics, such as maxima or inflection points, which are absent in the data. For example, it is easy to verify that by increasing the of the polynomial, the resulting curve tends to become more and more distant from the data. In practice, it is rare for a polynomial of degree greater than two to prove useful. However, it is often possible to apply transformations to the data in order to obtain polynomial laws.

Consideration of the resulting computation time also makes it clear that such methods are difficult to apply in 'real time', i.e. for 'on-line' identification of the parameters of a system. For on-line optimisation purposes, i.e. for automatic adjustment of the control parameters as a function of the variation of the actual system parameters, recursive forms

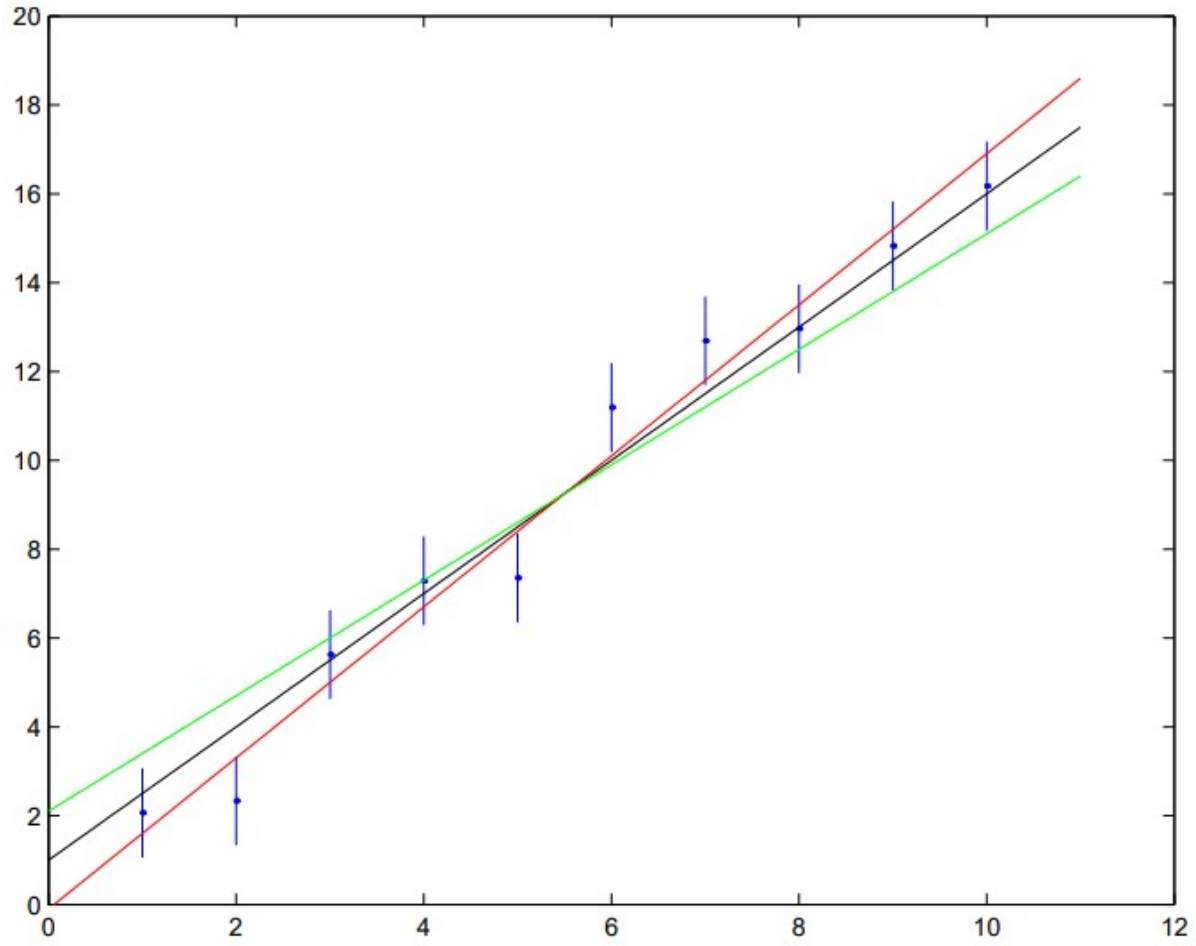

Figure 2.5: Linear fitting (correlation between intercept and slope)

(**third method**) of evaluation must be used which reduce the amount of calculation. The most commonly used search methods are the Gauss-Newton direction, the Levenberg-Marquardt and the steepest descent gradient search method ([7]), which are based on iteratively updating the parameters to be identified in such a way as to get closer and closer to their actual value. I will not expand on the description of this type of algorithm because in the thesis work we used a Mathlab function, called *greyest*, which at each iteration *greyest* chooses the search method to obtain the highest reduction in error.

2.3.5. Analysis of the residual

The identification of the grey-box model provides a more or less accurate model depending on the complexity and the amount of physical knowledge brought into the model. Under any circumstances, the model based on first principles will not be able to perfectly mimic the real system as it has a lower order structure. The difference between the real output value and the simulated value (provided by the identified model) is called residual or error

and contains parametric uncertainties, higher-order dynamics, noise, etc.

$$error = y_{real} - y_{grey-box}$$

The next step in the procedure is to analyse the residual and to assess whether a linear model of higher-order tends to fit the residual dynamics. We will limit ourselves to the use of linear structures since it is not possible to identify what kind of non-linear relationship the error system presents. In general, the error system will exhibit non-linear behaviour, but being "small" compared to the behaviour of the real system (main dynamics), the non-linearities can be considered as disturbances. Therefore the implementation of a linear model has a high probability of fitting the residual system. As mentioned above, we will limit ourselves to using ARX structures that have very good potential for identifying an (approximate) linear system.

However, if one wants to obtain a model that is well matched to the physical system, one must focus on a non-linear model. Dealing with a generic nonlinear dynamical system is a task of great difficulty, so we usually restrict ourselves to considering specific classes of nonlinear systems. One such class is that of block systems, which are characterised by particular sequences of static nonlinear blocks (in the sense that their output at a certain time is uniquely determined by the value of their input at that time) and linear dynamic blocks, which are cascaded. This type of model approximates non-linear physical systems quite well and is also particularly well suited to the automation sector since with the various blocks it is possible to separately schematise the non-linearities due to actuators, sensors, plant, and so on. A problem with all block systems is that, in general, there are infinite sets of parameters that provide the same input/output relationship for the system.

The main types of non-linear dynamic block systems are Hammerstein-type systems, Wiener-type systems and mixed Hammerstein-Wiener and Wiener-Hammerstein systems. In particular, we will implement a block system in the first case study, since the analysis of the dynamics of the grey-box model and the corresponding error system show a static correlation between input and output.

The investigation of more complex empirical methods is an inefficient choice because the model that estimates the error contributes only a small part. Therefore, we do not expect high accuracy in error modelling as the residual dynamics are only high order dynamics mixed with stochastic dynamics that we do not care to fit. The grey-box model will have done the bulk of the work by identifying the main features of the system and physically motivating them in such a way that they are interpretable.

Subsequently, using the MATLAB System Identification toolbox [9], it is possible to create a structure that combines all possible model-order combinations and calculates the loss function for each combination by estimating the ARX models. The model-order combination, among those considered, that best fits the validation data will be searched through the application of the criteria introduced at the beginning of the chapter (FPE, AIC). The model with the lowest loss function values is selected and used for the final estimation of the ARX parameters.

2.3.6. Evaluation of the results

Finally the procedure ends by combining the simulated response of the Grey-box model and the simulated response of the ARX model creating a parallel structure. The last step is the acknowledgement and evaluation of the results indicated by the chosen figure of merit (in our case NRMSE/fit). As will be exemplified in the case studies, the contribution of an empirical model may be superfluous or even unnecessary because the gain produced is irrelevant. We can distinguish two scenarios:

- the empirical model has made a significant improvement on the first-principles model, so through a hybrid approach the grey box model has gained accuracy.
- the black-box model did not improve the response of the grey-box model so it can also be discarded and the use of a pure grey-box model is preferred.

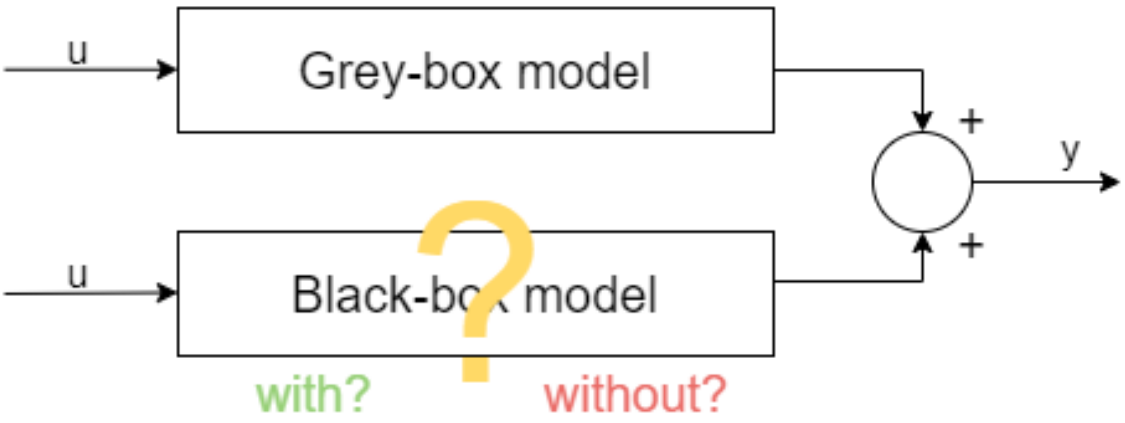

Figure 2.6: Is the introduction of a machine learning model relevant?

The last scenario is a counter-evidence of what physics had provided during the construction of the grey-box model. In other words, the irrelevance of adding a stochastic module within the identification procedure is synonymous with the holding of low authority.

The choice of imposing a higher or lower authority on the first-principles model was carried out before making an identification of any kind (white, grey or black) as it must adhere to the physics of the system and the purpose of the implementation. Thus, the validity of using a hybrid system can be deduced before acting with the identification because based on how the grey-box model will be formed we will already know if the residual dynamics will be significant (it is the physics that conveys this and not the data set).

The opposite procedure to the one provided by this thesis is to not implement a physically motivated model and to use linear, non-linear or stochastic fitting techniques for the identification of the actual process. However, this approach is totally related to the type of black box architecture we implement and the fact that the decision of a particular approach is not physically motivated makes the procedure restrictive. It does not allow us both to make conscious decisions in itinere and to physically justify step by step the results obtained.

A scenario that we will not evaluate experimentally but may arise is the possibility that the addition of a black box model may worsen the predictions obtained from the grey-box model. The origin of this circumstance is the poor ability of a data-driven model to operate in non-canonical or unexplored regions in the training phase. A grey-box model does not have this problem because it is physically motivated so it will always produce a response that is interpretable. The physical value of a grey-box model is an enormous advantage because the response only describes physically justifiable correlations and excludes the introduction of spurious correlations that an empirical model might generate. In order to verify this circumstance, the excitation of the two models through a signal which, although physically justified, explores a region far from the canonical working condition, is required.

3 | Case Study

In this chapter, the proposed procedure is applied to two simulated real processes which differ in their characteristics in the time domain and in the frequency domain. In each case, the chosen processes lie within the sphere of interest of the research group (the energy and electrical area) so as to consider applications in which understanding is maximised.

3.1. Case study 1

3.1.1. Real Model

The first model concerns the thermodynamic study of a rod. We created a detailed model of a rod with an elevated number of variables to better mimic its heat dynamics. The heat transfer phenomenon is solved by using a three dimensional lumped model that takes into account both convection and conduction phenomena. This model is composed of twenty lumps where each lump identifies a different region axially. Each of them has been introduced in the model as a material node representing the surface temperature. The first node has been placed near the source of power on the left and the last node is in touch with the external environment so the temperature of this node is equal to the external temperature.

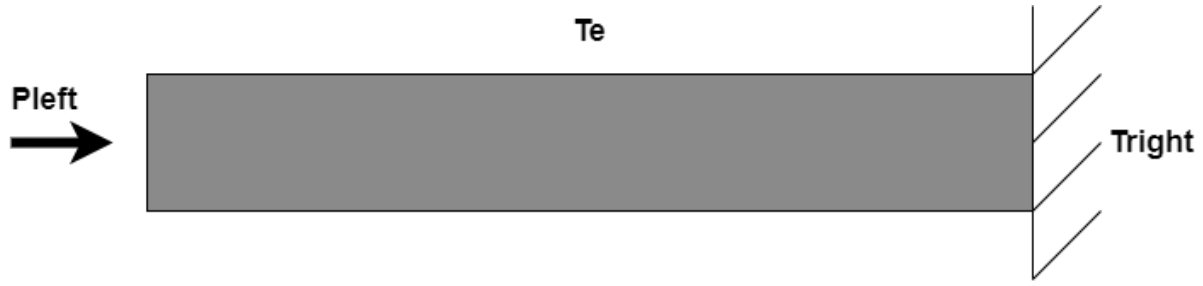

Figure 3.1: Rod illustration

This model is based on the electrical analogy where the conduction phenomenon is represented by resistance and the accumulation or release of thermal energy during a transient evolution is modelled by capacitors like the figure 3.2 shows.

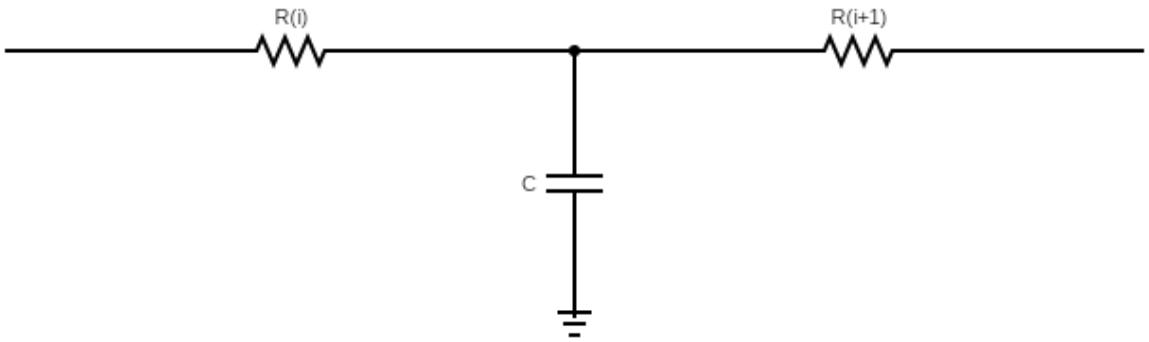

Figure 3.2: Electrical analogy of a rod lump

The proposed methodology is explicative if the model presents a suitable level of complexity otherwise the usage of a simpler identification approach is recommended. Therefore the simulated real system introduces some plausible details that a simple rod exhibits when it is located in a real environment. The first one interrogates the conduction phenomenon. Conduction is a process of heat transfer through solids. When a temperature gradient exists in a body, experience has shown that there is a transfer of heat from the high-temperature region to the low-temperature region. The heat transfer rate per unit area is proportional to the temperature gradient given by:

$$\frac{Q}{A_{cross}} \propto \frac{\Delta T}{\Delta X}$$

Where Q is the heat transfer rate, A is the cross area of heat transfer, $\frac{\Delta T}{\Delta X}$ is the temperature gradient in the direction of heat flow. When the proportionality factor is inserted, we get:

$$\frac{Q}{A_{cross}} = -K\frac{\Delta T}{\Delta X}$$

The positive factor K is called the coefficient of thermal conductivity of the material. The negative sign indicates that heat transfer takes place in the direction of decreasing temperature. The thermal conductivity coefficient is a physical property of the material. Although it is fairly constant in a narrow temperature range, it varies over a wide temperature range. In this study, the dependency of the thermal conductivity coefficient with the temperature is modelled as a linear relationship. The knowledge of the coefficient is limited to the first and the last lump, the middle lumps own a conductivity based on their location with respect to the interpolation line that passes through the two extremities.

The second one is related to the transient convection profile. Generally, the convective

heat transfer is described by Newton's law of cooling that defines a linear relationship between a heat flow and its corresponding driving force, δT, which is defined as the temperature difference directly on the wall, T, compared to the vicinity, T_e, multiplied by a characteristic parameter called the heat transfer coefficient, γ.

$$Q = \gamma(T - T_e)$$

Since this thesis is not a treatment aimed at the analysis of the heat transfer of heated rods but on the use of the black box model and grey box model, and therefore the model is for illustrative purposes, we introduce a non-linear dependence between the temperature variation and the γ coefficient that accounts for the different conditions of air motion or any other effect that affects the heat transfer. No specific correlation has been investigated in the current literature but the non-linearity allows us to consider a more complex model assuming unusual convection motion.

$$Q = \gamma(T - T_e)^{\alpha}$$

3.1.2. Data Analysis

In this section, the main topic is to describe the way we extrapolate the time series data and some considerations about why we decide to use a particular excitatory signal.

Input signal perturbation quality determines the effective variation in the system response, different input signal qualities have been compared. The random input signal to the process procures the whole dynamics of the process around the desired operating region. Capturing system dynamics using PBRS is more efficient since it comprises both positive and negative changes within the input sequence.

Furthermore, since the application operates in a wide range of conditions, we introduce a carrier signal made up of multiple steps of different sizes. The PRBS is incorporated in the carrier signal and it acts as a small white noise (3.3). The completeness of the signal created will allow us to identify models representing the order of detail that a common controller requires. Seeking greater complexity of the input signal would cause the identification tool to require more computational power. It will try to identify all the dynamics present in the data set, introducing high-order dynamics that are not significant to represent the process. This would lead to a probable over-fitting of features that are specific to the experiment itself, decreasing the ability to generalise.

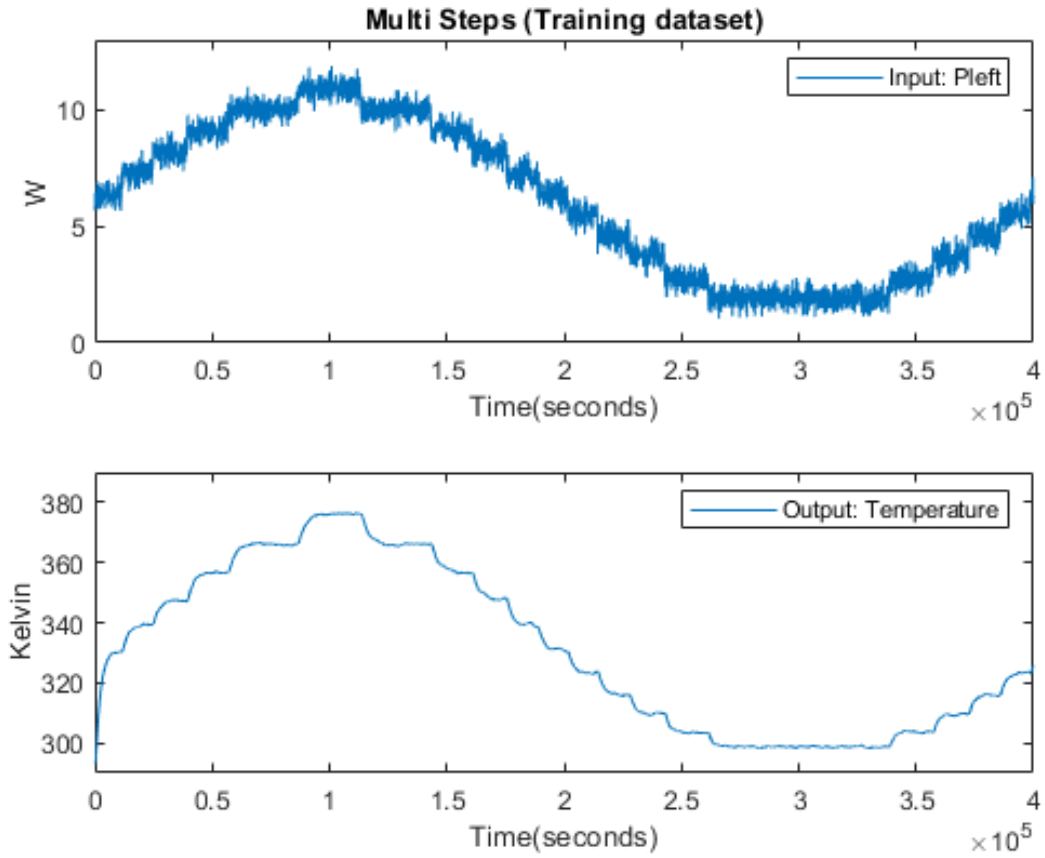

Figure 3.3: Multi steps signal as a training data-set

Secondly, to avoid the model performing extremely well on the samples used for training but performing poorly on new unknown samples, it is fundamental to consider a different type of experiment. The evaluation of the quality of the model is declared by an appropriate balance between the accuracy achieved by the model in the two data-sets. The training set is used to build the model with multiple model parameter settings and then each trained model is challenged with the validation set. The validation set contains samples with known provenance, but these classifications are not known to model, therefore, predictions on the validation set allow the operator to assess model accuracy. Based on general guidelines, the adoption of a step signal is advised because it provides interpretable insight into the settling time, the rise time, the presence of overshoot and offset; in short, all of the qualities related to the model time behaviour.(3.4)

However, the quality of the model cannot be verified by data sets used in the model tuning procedure, so it is necessary to carry out an additional experiment of a different nature. The data collected during the experiment are called test data-sets and allow us to check whether the identified model is able to operate in different working regions than

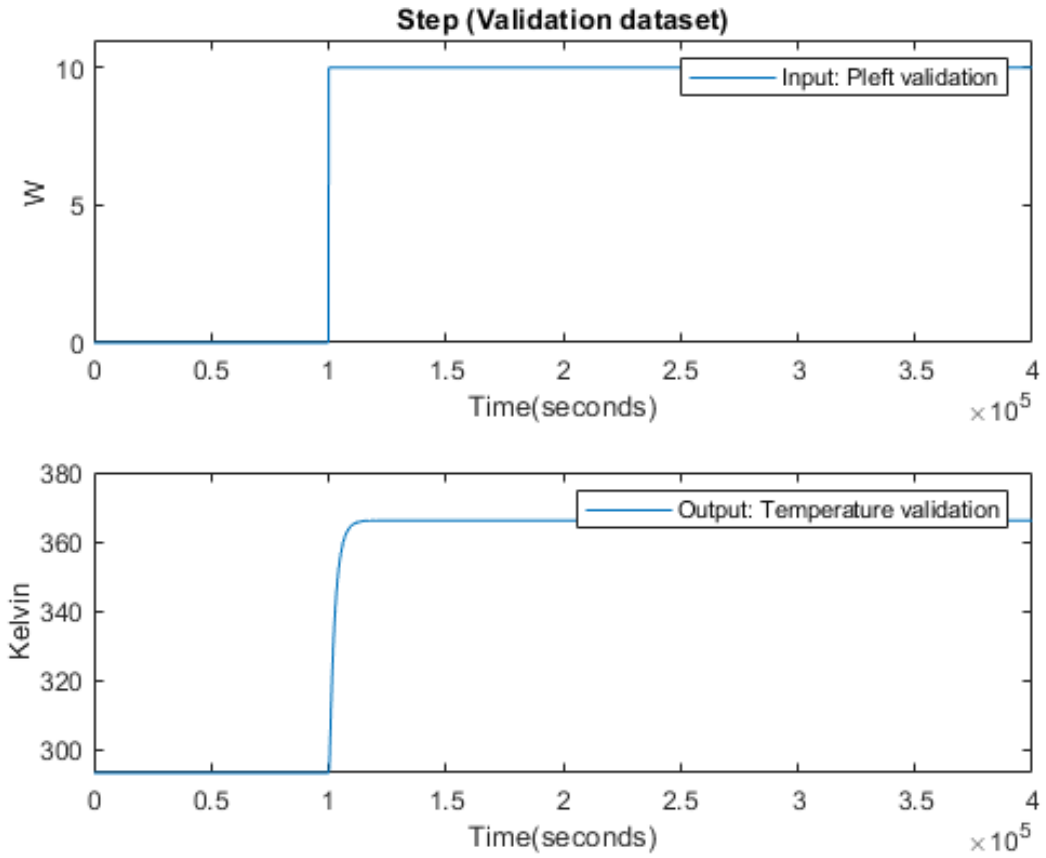

Figure 3.4: Step signal as a validation data-set

those explored so far. It is essential that the model is accurate in this scenario because when it is implemented in reality it will have to perform in a wider working region. The ramp input signal is used to test the performance of the model in different scenarios with respect to those analyzed before.(3.5)

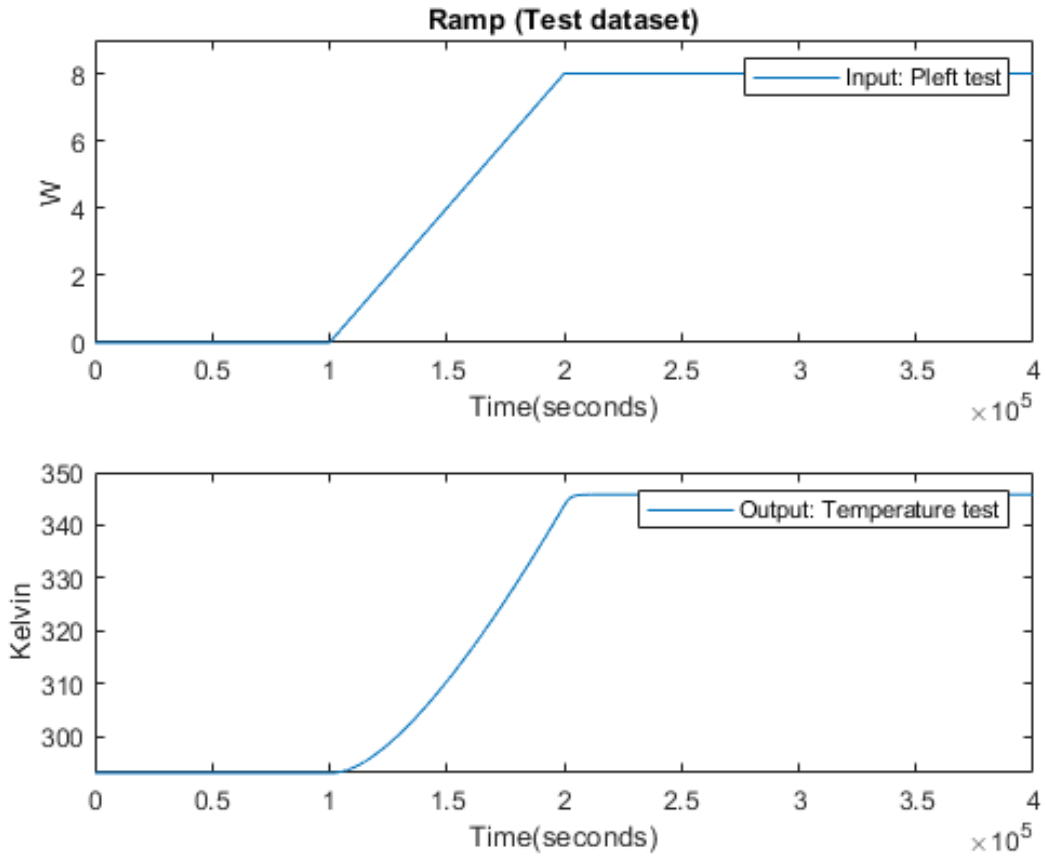

Figure 3.5: Ramp signal as a test data-set

3.1.3. Hybrid Model

In this chapter, the grey-box modelling approach will be applied to the rod and the heat transfer coefficient will be identified. As explained in the Chapter 2, this kind of modelling usually requires combining partial a priori knowledge of the object of consideration with empirically obtained data. However, in the lack of real data, the data obtained from a Modelica model could be used instead. Since the rod is an imaginary model created for research purposes (not a real application), measurements and collection of the empirical data are not feasible. Accordingly, the data needed for the parameters identification process in this work is obtained from the white-box simulation in Modelica. The modelling of the grey-box model is carried out for three hypothetical purposes:

1. tuning of a controller;
2. approximate simulation;
3. simulation.

In practice, it results in the identification of three first-principle models with different levels of detail and complexity.

Grey-box model for tuning a controller plus ARX

The purpose of the first grey-box model will be to tune a controller (let's consider a standard PID for convenience). The tuning procedure requires that the plant P is approximated by a first-order system.

$$P(s) = \frac{\mu}{1 + s\tau}$$

where μ is the gain and τ is the time constant.

The parametric identification is carried out through the method of the areas that through a recording of the step response (previously presented in Fig. 3.4) of the process gives the value of mu and tau that we will indicate in the table 3.1.

mu	7.2973
τ	2522.01

Table 3.1: Identified mu and τ by method of areas

Next, we simulate the time response of the first-order model against a "complete" signal introduced previously in Fig 3.3 The time response is very approximate and manages to mimic the essential dynamics of the system. The percentage of fit of the grey-box model is very low but it was predictable and compatible with the purpose of the identified model: it was not created to have to fit all the dynamics of the system but rather to have a structure compatible with that required by the purpose. Fig.3.6

The time response is very approximate and manages to mimic the essential dynamics of the system. The fit rate of the grey-box model is mediocre but was predictable and compatible with the purpose of the model identified: it was not created to fit all the all deterministic and stochastic relationships of the system but rather to have a structure compatible with that required by the purpose. The occurrence of a non-linear static correspondence can be seen through the error representation shown in Fig.3.7, where the signal is not comparable to white noise, i.e. a signal with zero mean.

More specifically, if I try to fit the existing relationship between input and error, I notice that the static correlation is pseudo-linear, at most quadratic as shown in the Fig.3.8.

At this point, the grey-box model has exhausted all its potential and the error between

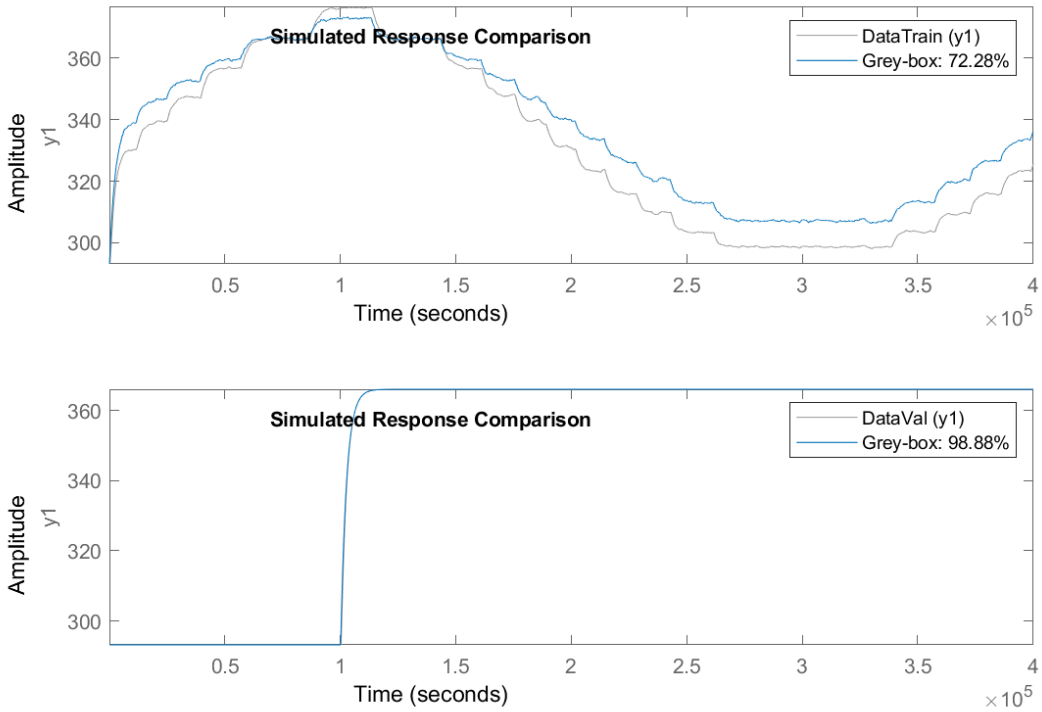

Figure 3.6: Time response of the first order grey-box model

the real and simulated model cannot be smoothed out using the first-principles model. Therefore, we will take advantage of a black-box model, in particular an ARX structure, in order to fit the residual dynamics.

The algorithm for selecting the ARX model relies on criteria such as AIC,FPE to decide on the best structure (see Chapter 2). The graphical representation of the model selection is described by Fig.3.9 which suggests the use of an ARX4411111 structure as the contribution of a higher order structure does not help to decrease the loss function.

The ARX model is linear in its parameters, so, given the correlation analysis between input and error, we expect that it will not be able to model all residual dynamics. The structure of the black box model is too simple to record acceptable values of fit as can be seen in Fig.3.10.

In fact, the injection of signals of a different nature into the hybrid model results in a deterioration of the model's accuracy compared to the response that the grey box model provides. The justification for this is due to the low priority of the grey box model compared to that acquired a posteriori by the black-box model. The grey-box model failed to guarantee generalisation of the procedure as the introduction of spurious correlations

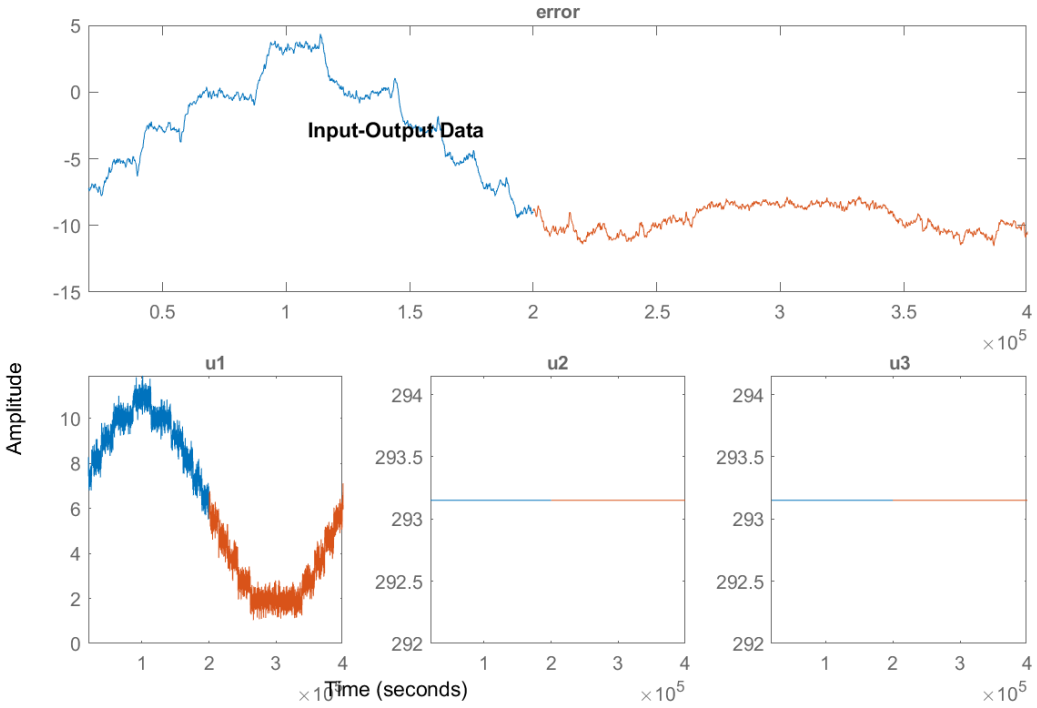

Figure 3.7: Error between the real process and the simulated one

Input	Grey-box fit%	Hybrid fit%
Multisteps+PRBS	72.28	88.49
Step	98.88	78.92
Ramp	75.45	72.29

Table 3.2: Percentage fitting values comparison

by the data-driven model was dominant. The graphs of the responses of the implemented models are avoided and replaced by table 3.2 which represents the percentage values of grey-box and hybrid fit against the input signals declared in Chapter 2.

This type of physical process cannot be accurately simulated by a first-order model because of noticeable non-linearities. For this reason, we decided to apply the procedure to the same case study but changing the purpose of the identified model: from PID tuning to approximate simulation. This will give us more freedom in choosing the structure/complexity of the black box model because we will no longer be limited to considering only first-order structures.

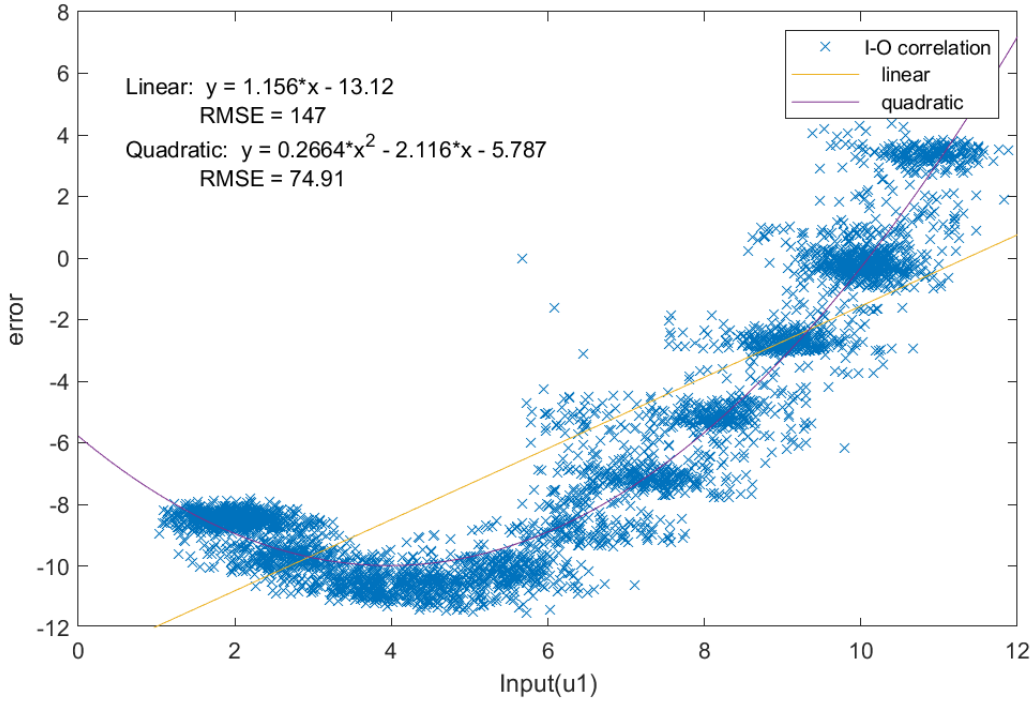

Figure 3.8: Scatter plot (Pleft vs Error)

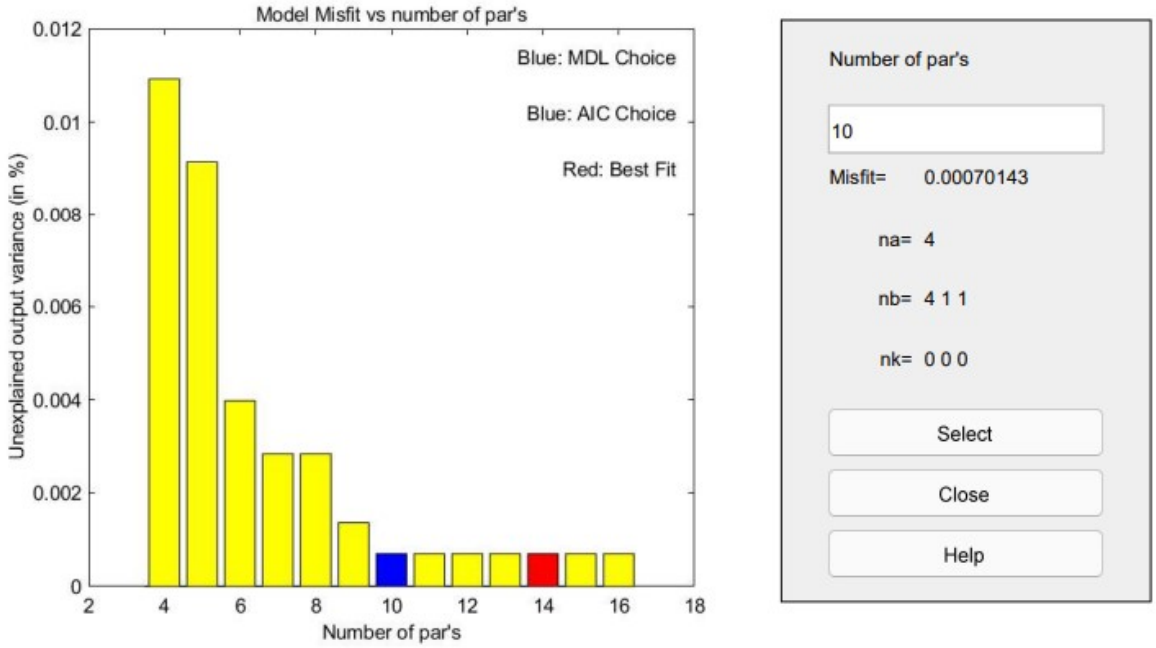

Figure 3.9: Proposed order for ARX model

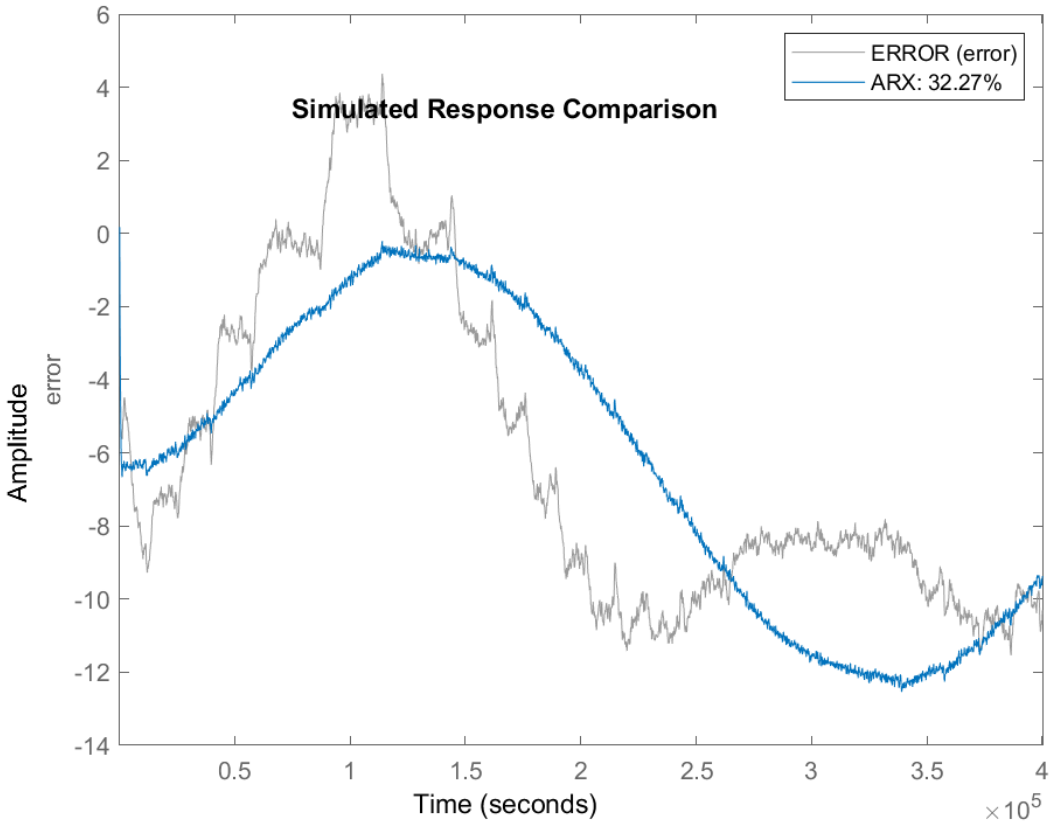

Figure 3.10: ARX model response

Grey-box model for approximate simulation plus ARX

The shift in the purpose of identification gives us the possibility to evaluate a more complex structure of the grey-box model where also the static correspondence relating input and output has been added directly into the model. Now the gain mu of the first-order model is no longer a constant value but depends linearly/quadratically on the input P_{left}.The dynamical system under consideration, shown in Fig.3.11, can be treated as a Hammerstein system since it consists of a static nonlinear block $f(u)$ and a dynamic linear block $P(s) = \frac{1}{1+s\tau}$.

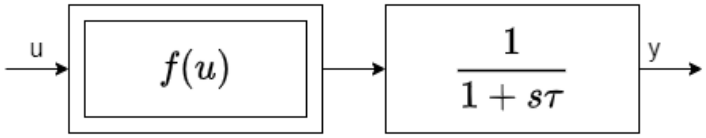

Figure 3.11: Dynamic system treated as Hammerstein

The function $f(u)$ has been implemented in order to compensate for a static correspondence in the variables that is evidenced by graphing the relation between input and output in stationary regime through a scatter plot in (Fig.3.12).

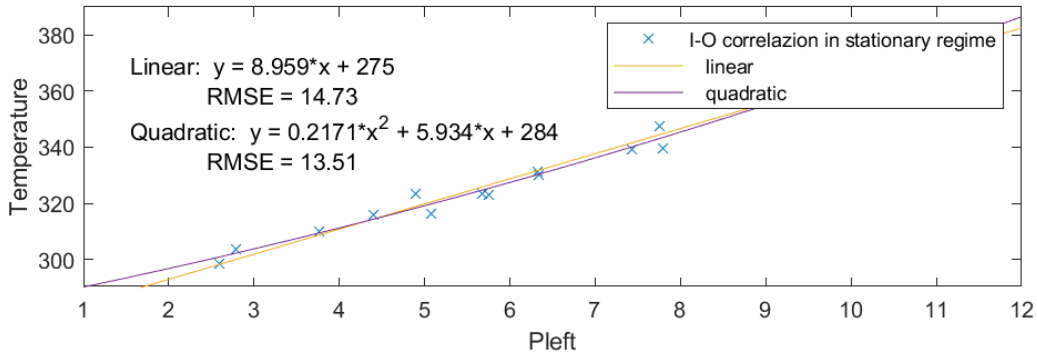

Figure 3.12: Scatter plot (Pleft vs Temperature)

The relationship between input and output can be represented by either a linear or a quadratic function as shown in Fig.3.12 We decided to use a quadratic compensation as the linear one was too weak to compensate for the static correlation.

The fitting percentage of the grey-box model is very high, but as can be seen from the auto-correlation plot of the residual (FIg.3.13), the confidence interval is not respected, which means that the correlation between input and error is statistically significant.

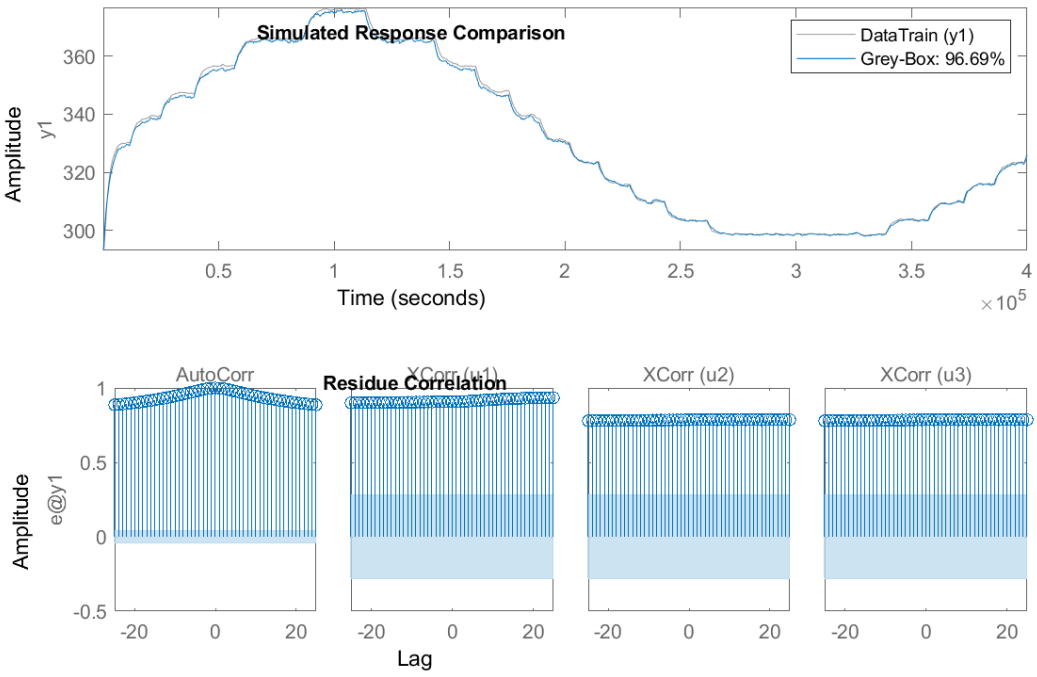

Figure 3.13: Simulated response and residue correlation plots of Grey-box model

As the error cannot be evaluated as white noise, the grey-box model is not sufficient to model all residual dynamics. It is the task of an ARX model to compensate for the limitations of such a structure by identifying higher order dynamics.

The structure chosen by the AIC and FPE criteria is ARX1411100 and as shown in Fig.3.14 the black-box model interpolated the error model well and identified the main characteristic. The value of archived fit% could be misleading but if we analyze the graph provided by the whiteness test we notice that the residual dynamics are almost comparable to white noise.

Finally, the hybrid parallel structure model is created and the interpolative properties of the black box model improved the accuracy of the grey-box model by identifying residual dynamics. For a simulation purpose the combination of two models with this complexity is enough since the fit rates improved in all 3 experiments as the table 3.3 summarises.

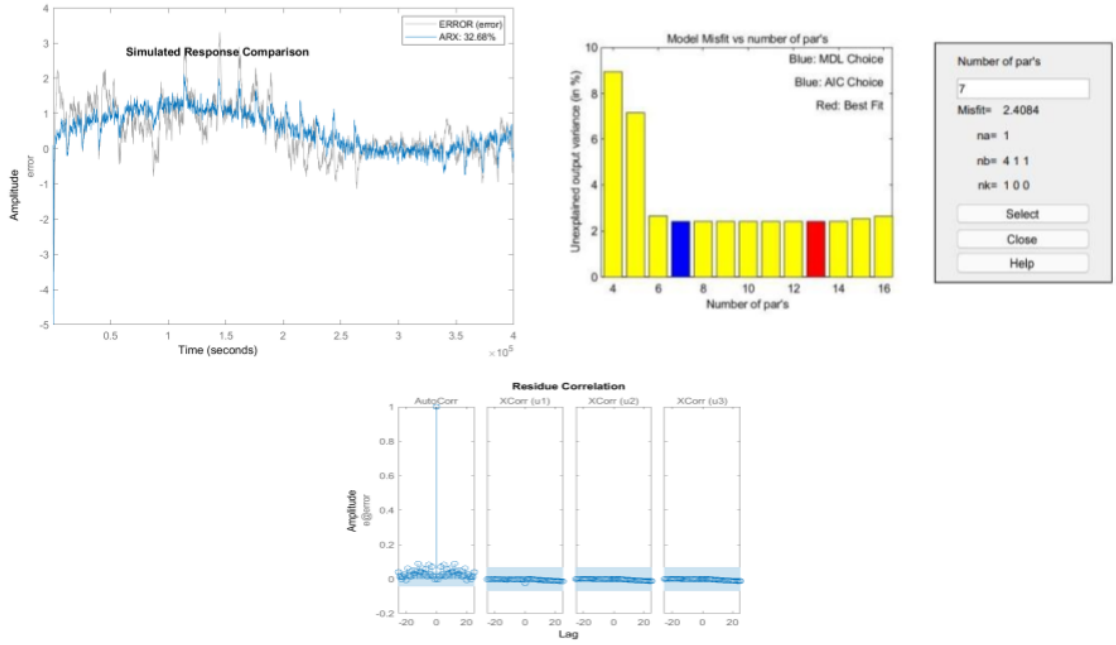

Figure 3.14: ARX1411100 performance

Input	Grey-box fit%	Hybrid fit%
Multisteps+PRBS	96.69	97.9
Step	96.75	99.15
Ramp	95.99	97.31

Table 3.3: Percentage fitting values comparison

Grey-box model for simulation plus ARX

The addition of a third variant of the approach that considers the development of a hybrid model to simulate the real process in detail could be avoided as the accuracy of the newly implemented model is already high. However, the implementation of a grey-box model with lumped-parameters is considered because it provides the user with more interpretability. With this typology, the correspondence between the identified parameter and the real parameter is direct, i.e. the parameters of the grey-box model are an agglomeration of the real parameters but have the same physical meaning.

Grey-box modelling implementation used for this purpose, in the form of the Mathlab function, was developed and provided by the undersigned. It is based on the thermal-electrical analogy, so the rod is modelled as a corresponding RC network.

According to this model, four main factors are affecting the heat transfer dynamics – the geometry of the rod, the thermal conductivity of the material, the heat capacity of the material and the coefficient of heat transfer with the environment. Moreover, the simplified model of the rod is composed of 2 lumps and every lump is designed as an electrical node. Additionally, the model integrates a heating system that provides heat through a thermal power source, while the external temperature and the last cross-section area are represented as thermal constants.

Before starting with the identification of heat transfer parameters several inputs are required to define the structure of the application: the length, the diameter, the density of the rod and the initial temperature of the lumps. They allow retrieving the cross-area, the mass and the lateral surface of the lump. Since the geometry of the rod is known, the listed parameters are not identified by the tool but they remain fixed throughout identification.

Therefore, the grey-box model is formed by the following equations 3.1

$$\begin{aligned} C\dot{T_1} &= P_{left} - \frac{4\lambda_1 A_{cross}(T_1 - T_2)}{L} - \frac{\gamma_{lat} A_{lat}(T_1 - T_e)^{\alpha}}{2} \\ C\dot{T_2} &= \frac{4\lambda_1\lambda_2 A_{cross}(T_1 - T_2)}{L(\lambda_1 + \lambda_2)} - \frac{4\lambda_2 A_{cross}(T_2 - T_{right})}{L} - \frac{\gamma_{lat} A_{lat}(T_2 - T_e)^{\alpha}}{2} \end{aligned} \tag{3.1}$$

Then, the parameters identification process is performed employing as an input signal the multi-step one because, as I described before, the adoption of a complete and well frequency distributed signal brings to the identification tool the amount of potential information needed.

The parametric identification was carried out by means of Matlab's *greyest* function which combines several search methods to obtain the best value for each uncertain parameter as already indicated in chapter 2.The Optimisation toolbox chose to use the Trust-Region Reflective Newton algorithm, which is based on the optimisation concept defined as a trust region. The main idea is to approximate a function $f(x)$ by a simpler function q that mimics the function f in a neighbourhood N around point x. The neighbourhood is called the *trust region*. The algorithm is iterative because the point x is updated at each step until it reaches convergence [7].

The identified uncertain parameters are summarised in the following table (3.4).

Parameter	**Initial value**	**Estimated value**
λ_1	70	61.06
λ_2	70	132.83
γ_{lat}	20	54.50
α	1	0.53
C	137.3	80.11

Table 3.4: Parameter estimation summary

The accuracy of the grey-box model structure is also reflected in a high degree of precision in fitting the dynamics of the actual process, as confirmed by the fit% in the table 3.5.

Input	**Grey-box fit%**
Multisteps+PRBS	98.27
Step	98.68
Ramp	75.93

Table 3.5: Percentage fitting value of the grey-box model

In this specific scenario we expect the residual dynamics to be modelled to be minimal and the contribution of an ARX structure to be minimal as it has the task of identifying only high-order dynamics comparable to noise. The same best order selection procedure for the ARX model was carried out and the results obtained are shown in Fig.3.15.

The overall model had higher accuracy in all three injected signals. The major difference of a grey-box model consisting of differential equations compared to a first-order model is the addition of interpretability in the model. Changing an estimated parameter within the physically motivated model results in an error that retains physical meaning. For the sake of clarity, let us show the percentages of fits stored in the table 3.6.

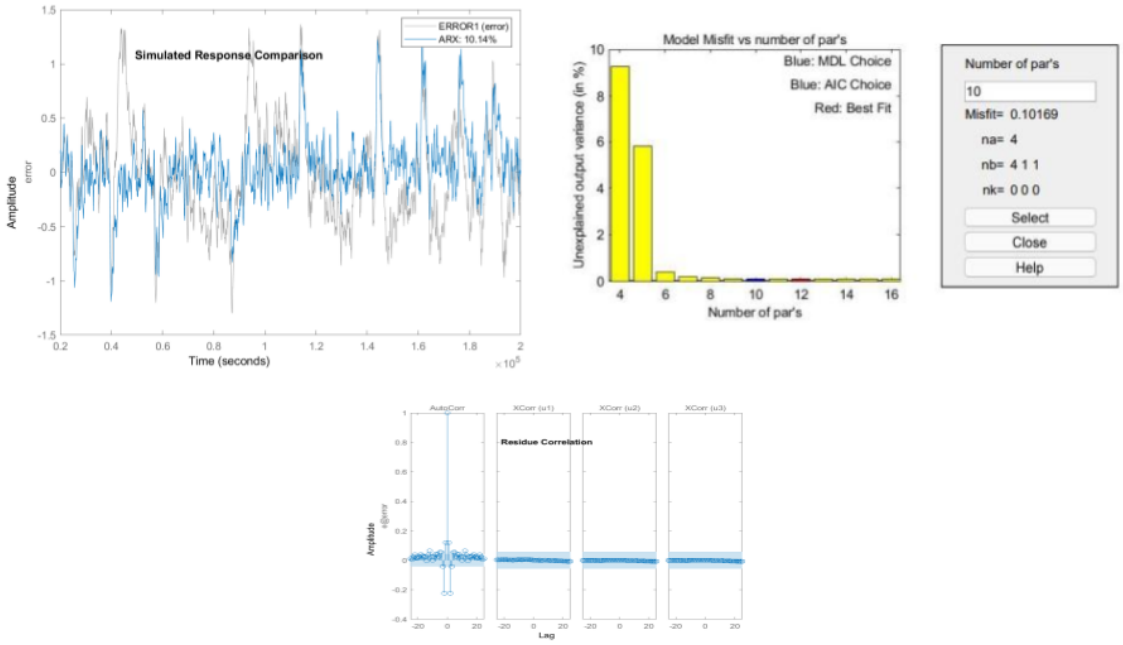

Figure 3.15: ARX4413000 performance

Input	Grey-box fit%	Hybrid fit%
Multisteps+PRBS	98.27	98.59
Step	98.68	98.87
Ramp	75.93	75.98

Table 3.6: Percentage fitting value comparison

3.2. Case study 2

3.2.1. Real Model

The second system consists of a set of 10 resistance-capacitance (RC) which will be modelled using 3 resistance-capacitance (RC) lumped element models. The resistors taken into account are the resettable fuses also called PPTC which, compared to the classical resistors, have a non-linear dynamics because they depend on the voltage. In other words, in the presence of a strong increase in voltage between the ends of the resistor and when the characteristic voltage is exceeded, the PPTC strongly lowers its resistance (in a non-linear way). The graphical representation of the second case study is shown in Fig.3.16.

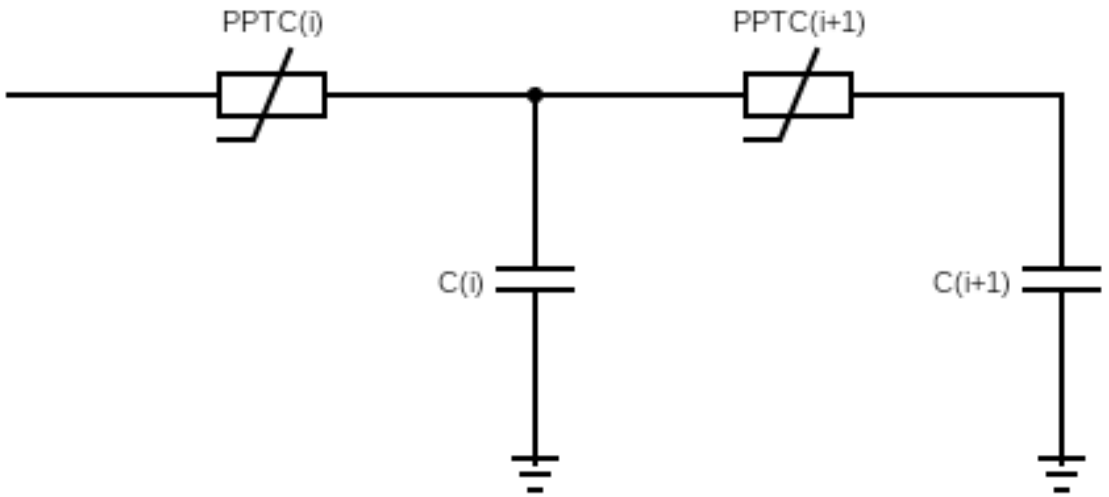

Figure 3.16: Case study 2 representation

The non-linear characteristic linking voltage and current is described in Fig.**??** where it can be seen that when the voltage applied to the varistor is below its threshold, the current through it is extremely small, which is equivalent to a resistor with infinite resistance. That is, when the voltage applied to it is below its threshold, it is equivalent to an off switch. When the voltage applied to the varistor exceeds its threshold, the current through it increases sharply, which is equivalent to an infinitely small resistance. In other words, when the voltage applied to it is above its threshold, it is equivalent to a closed switch.

The parameters of the grey-box system agglomerate the parameters of the real system into fewer units in order to have a model of lower complexity. For the sake of completeness,

we have decided to integrate this case study into the thesis because there is a noticeable difference in the frequency domain between the real system and the first-principles model. In other words, if we visualise the response of the grey-box model in the time domain it is reliable in capturing the main dynamics, but in the frequency domain the error is considerable. The following feature can be seen if we inject into the real process a chirp-type signal in which the instantaneous frequency grows linearly with time.

3.2.2. Data Analysis

The tuning, validation and testing of the hybrid model requires the implementation of three different experiments in order to obtain three data sets simulating the response of the real system at a suitable range of operating conditions. In this particular case study, the following signals were used to carry out the three main steps of our proposed procedure:

- a step for the tuning of the grey-box model;
- a chirp signal for the detection and identification of the residue;
- a square chirp signal for the test.

Their graphical representation is shown in Fig.3.17.

3.2.3. Hybrid model

The case under consideration is implemented in Modelica as a white-box model and will be considered by us as a hypothetical real process. The purpose of the identification is purely simulative as the addition of this case study has the task of testing our proposed approach to a process with different characteristics. The introduction of a first-principles model is a solid basis on which to establish a more accurate identification defined by the black-box model.

The grey-box model consists of three resistance-capacitance modules, as previously announced, where the resistances and capacities attempt to mimic the dynamics of the ten modules that make up the real system. The fit of the physically motivated model will not be optimal as each module of the ten provides a non-linear contribution to the overall dynamics through the presence of Varistor resistors. Specifically, the capacitors are linear capacitors whereas the resistors are non-linear and are described by the following equation :

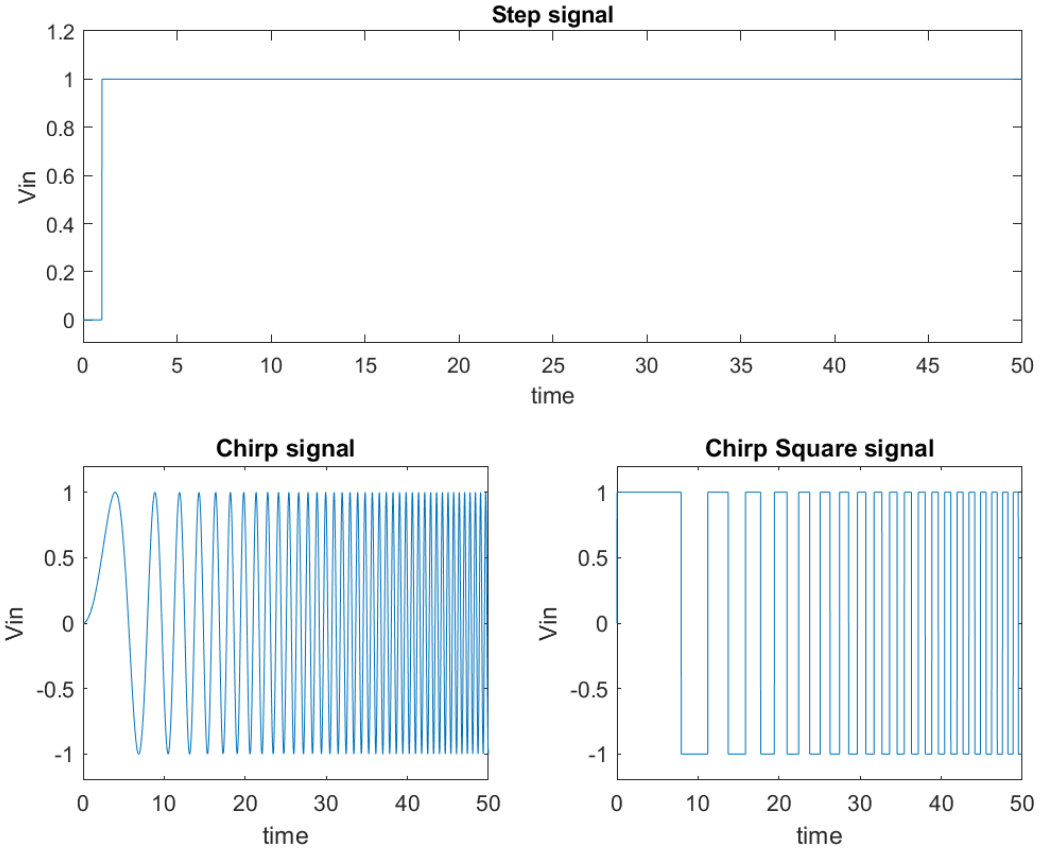

Figure 3.17: Three experiment data-sets

$$R = R_0 + K_{ldr}|i|$$

where the resistance varies linearly (K_{ldr}) with the modulus of the current (V-I characteristic is symmetrical to the origin).

The identification of the parameters was carried out by means of the second method introduced in Chapter 2, i.e. by interpolation of the points through lines and polynomial curves. The aim of the identification is to minimise the "distance" between the real and estimated points in order to create a curve that best approximates the real process. In the specific case, the values of the three resistances and the three capacitances are shown in the following table(3.7).

The model was identified using a step signal and as depicted in Fig.3.18 the model appears to mimic the real process with high accuracy but if we look in detail the suspicion that in the frequency domain, the accuracy will decrease is high. In fact if we inject to the

Parameter	Estimated value	Parameter	Estimated value
R_{01}	4e4	K_{ldr1}	1e4
R_{02}	2e4	K_{ldr2}	1e4
R_{03}	6e4	K_{ldr3}	1e4
C_1	5e-6		
C_2	2e-6		
C_3	6e-6		

Table 3.7: Estimated value of the parameters

system a chirp signal we notice that the residual present between the identified model and the real model increases linearly with the increase of the frequency so the more the input signal has high frequencies the more the grey-box model will not be able to mimic the dynamics of the real system.

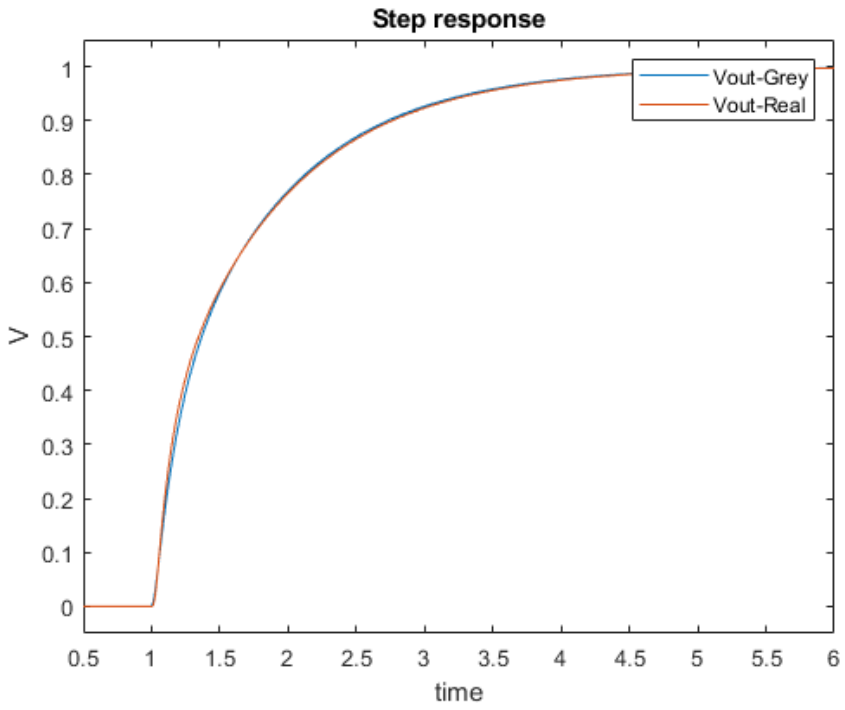

Figure 3.18: Step response comparison

Therefore, the use of an ARX model is likely to identify residual dynamics. The choice of the most efficient structure was made using the *selstruc* function of Matlab and settled on an ARX343 which mimics the error system with high accuracy as can be seen in Fig.3.19.

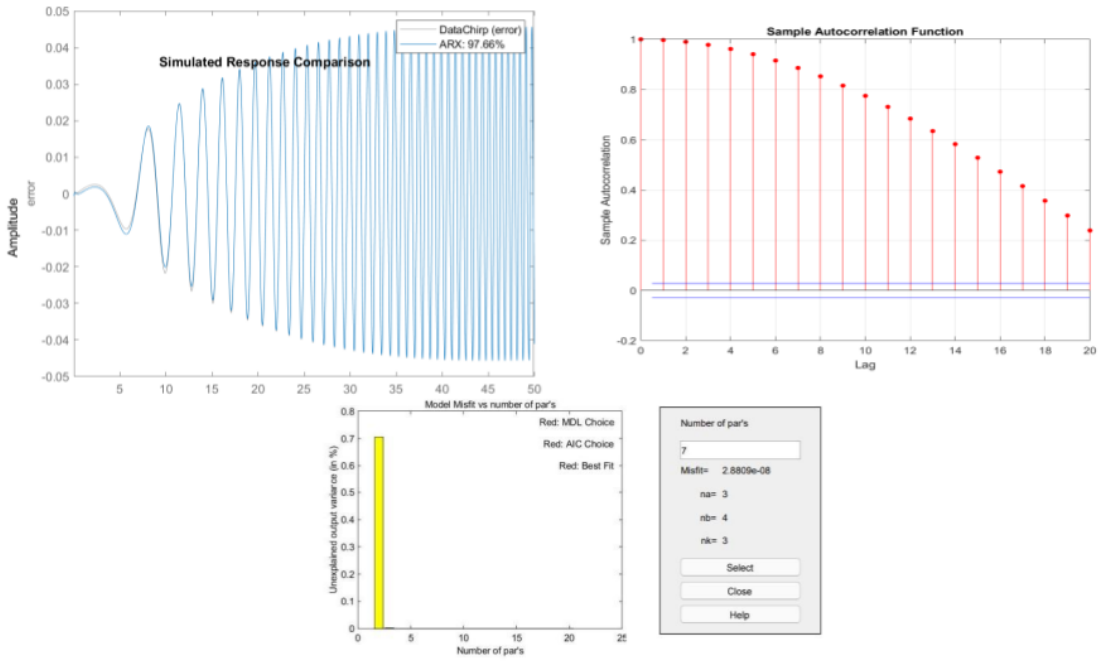

Figure 3.19: Identification of the residual dynamics by ARX343

The final test concerns the performance comparison between the implemented grey-box model and the parallel structure hybrid model.

The test was done by injecting a different signal from those used for the tuning of the two models to ensure the generalisability of the hybrid model. Therefore, against a Chirp Square signal the integration of the ARX model improves the grey-box model. The actual improvement in accuracy is not very visible in the time domain but if we compare the fit rates, the increase is substantial (see Fig.3.20).

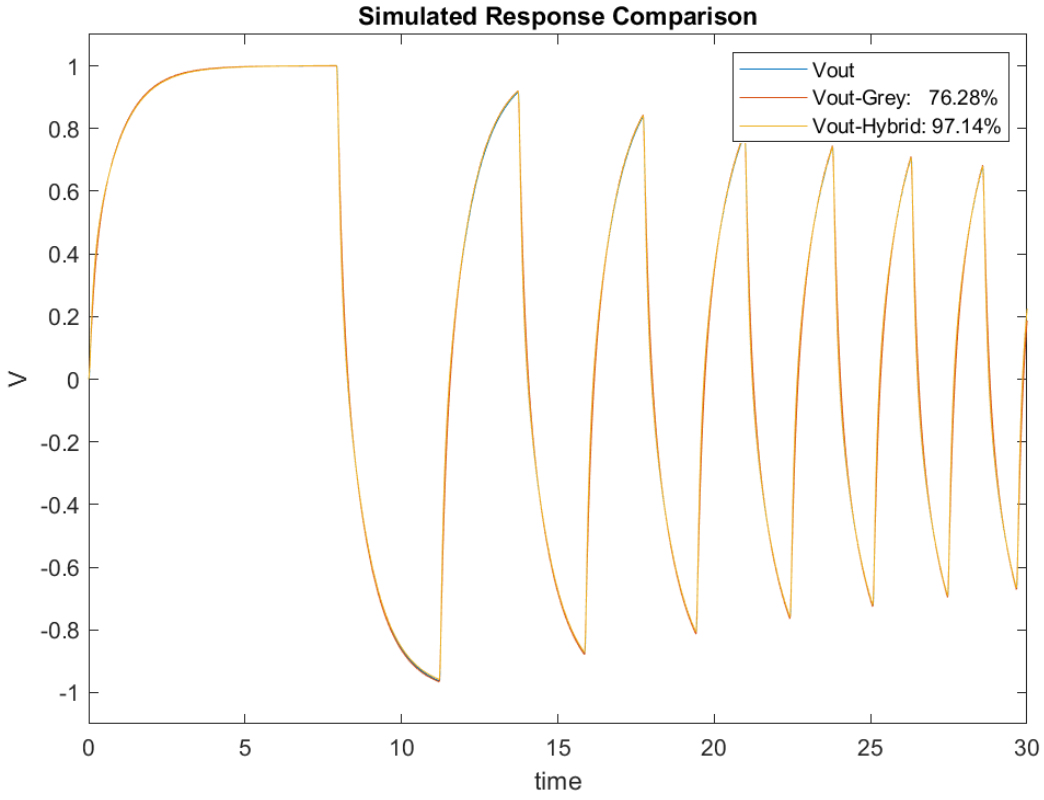

Figure 3.20: Chirp Square response comparison

4 | Conclusion and Future work

The aim of this thesis was to propose and apply a revised approach of the hybrid identification method. The reason why the proposed method is advantageous over those already found in the literature is its extreme versatility and its ability to combine the merits of a first-principles model and a data-driven model.

The flexibility of the proposed methodology makes it possible to adapt the identified model to a variety of purposes from the tuning of a controller to a more or less detailed simulation of the process under consideration. The possibility of adapting the structure of the hybrid model according to the purpose and the physics governing the process we have to identify is given by the concept of variable authority. This means that the authority of the model of principles acquires a priori the authority that allows it to adhere to the purpose for which it is implemented and in the same way to be faithful to the physics of the process. In retrospect, the black-box model will acquire the remaining authority supporting the potential acquired of the grey-box model.

The desire to combine the two approaches in a parallel structure stems from the need to create a model that is both interpretable and accurate at the same time.

The first quality is guaranteed by a model that can be constructed using differential equations that possess parameters with a high physical meaning. It also has the potential to be extremely generalisable without being constrained by limited working conditions that would make it inflexible and sometimes inconsistent with the dynamics of the process.

The extreme accuracy is guaranteed by the implementation of a data-based model to which the physical context it has to identify is obscure, but it has strong interpolation capabilities. This aspect is reflected in applications with complex dynamics that are at best non-linear and difficult to analyse using physical equations. The limited potential that a physical model possesses in these types of scenarios is overcome by the stochastic nature of a black-box model which, by referring only to the data, manages to capture all the high-order dynamics that would be impossible to incorporate in a grey-box model.

The methodology has been applied to two case studies with different characteristics in

the time and frequency domains in order to provide a complete picture, also analysing borderline cases in which the implementation of a black-box model with high acquired authority can lead to disadvantages in the replication of a real process as it introduces spurious correlations (not physically explainable).

Possible future work could be summarised in five main points:

- Firstly, expanding the case studies in which the methodology has been implemented beyond those related to thermodynamics or the electrical field.
- Secondly, to tune a controller and assess whether the use of a grey-box model implemented in this way works in practice. Above all, assess the response of the controller to the subsequent introduction of the error model to the overall model.
- Thirdly, to use this methodology to introduce robust synthesis techniques. In other words, I define the nominal model the first-order model used to calibrate the controller, then I identify the residual through a black-box model, finally I use the data-based model to provide an estimate of the model error as the controller would perceive it and make the controller robust to at least this model error.
- Fourth, provide a possible formal development on a subclass of real nonlinear systems by assessing the feasibility of residual model identification and achievable accuracy.
- Finally, apply the procedure in a real application case where the measured data are real and thus incorporate uncertainties such as random, systematic, stochastic errors

Bibliography

[1] S. M. H. Christian Paraiso Salah El-Dine and H. Werner. Black-box versus grey-box lpv identification to control a mechanical system. pages 5152–5157, Maui, Hawaii, USA, 12 2012. 51st IEEE Conference on Decision and Control.

[2] S. G. Christy Green. Residential microgrid optimization using grey-box and black-box modeling methods. *Elsevier Ltd*, Energy & Buildings 235:1–14, 2021.

[3] M. J. Ellis. Machine learning enhanced grey-box modeling for building thermal modeling. pages 1–6, New Orleans, USA, 5 2021. American Control Conference (ACC).

[4] M. S. Francesco Massa Graya. A hybrid approach to thermal building modelling using a combination of gaussian processes and grey-box models. *Elsevier Ltd*, 165: 56–63, 2018.

[5] D. H. I. Skuliber and S. Dešić. Black-box and gray-box components as elements for performance prediction in telecommunications system. page IEEE Xplore, IEEE Xplore, 2009. ConTEL 2009. 10th International Conference.

[6] O. N. Jan-Philipp Roche, Jens Friebe. Machine learning for grey box modelling of electrical components for circuit- and emc-simulation. pages 1208–1216, Germany, 7 2020. PCIM Europe digital days.

[7] M. J. Kochenderfer and T. A. Wheeler. *Algorithms for Optimization*, volume 520. MIT Press Ltd, 2019. ISBN 9780262039420.

[8] M. Krishnan. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technologys*, pages 489–493, 2019.

[9] L. Ljung. System identification toolbox (user's guide). *The Mathworks , Inc.,Natick,*.

[10] J. P. S. F. d. A. Moritz von Stosch, Rui Oliveira. Hybrid semi-parametric modeling in process systems engineering: Past, present and future. *Computers & Chemical Engineering,60*, pages 86–101, 2014.

[11] A. J. Qiang Xiong. Grey-box modelling and control of chemical processes. *Pergamon*, pages 1027–1039, 2002.

[12] S. R. F. T. F. G. Riccardo Guidotti, Anna Monreale and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51,5:2–10, 2019.

[13] B. D. K. J. L. S. Estrada-Flores, I. Merts. Development and validation of "grey-box" models for refrigeration applications: a review of key concepts. *International Journal of Refrigeration, 29*, pages 931–946, 2006.

[14] B. N. Sofia Rachad and B. Bensassi. System identification of inventory system using arx and armax models. *International Journal of Control and Automation*, 8,12:283–294, 215.

[15] B. Sohlberg. Hybrid grey box modelling of a pickling process. *Control Engineering Practice 13*, pages 1093–1102, 2005.

[16] J. E. Sohlberg B. Grey box modelling – branches and experiences. pages 1–6, Seoul, Korea, 11 2008. IFAC.

[17] J. J. Timur Bikmukhametov. Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models. *Elsevier Ltd*, pages 1–27, 2020.

[18] J. P. W. L. Xia Hu, Lingyang Chu and J. Bian. Model complexity of deep learning: A survey. *arXiv*, pages 1–36, 2013.

[19] M. C. Y. L. W. Z. Z.F. Wu, Jin Li. On membership of black-box or white-box of artificial neural network models. pages 1400–1403. ICIEA, 2016.

List of Figures

List of Tables

Acknowledgements

I wish to thank my parents for their moral and financial support.

I would like to thank Professor Alberto Leva for giving me the opportunity to carry out this thesis work and for always being available for any clarification.

I am also thankful to Professor Simone Formentin for having contributed to the thesis work by being willing to provide explanations.