Combustion Theory and Modelling

ISSN: 1364-7830 (Print) 1741-3559 (Online) Journal homepage: http://www.tandfonline.com/loi/tctm20

Numerically accurate computational techniques for optimal estimator analyses of multi-parameter models

Lukas Berger, Konstantin Kleinheinz, Antonio Attili, Fabrizio Bisetti, Heinz Pitsch & Michael E. Mueller

To cite this article: Lukas Berger, Konstantin Kleinheinz, Antonio Attili, Fabrizio Bisetti, Heinz Pitsch & Michael E. Mueller (2018) Numerically accurate computational techniques for optimal estimator analyses of multi-parameter models, Combustion Theory and Modelling, 22:3, 480-504, DOI: 10.1080/13647830.2018.1424353

To link to this article: https://doi.org/10.1080/13647830.2018.1424353

Published online: 06 Feb 2018.

Combustion Theory and Modelling, 2018
Vol. 22, No. 3, 480–504, https://doi.org/10.1080/13647830.2018.1424353

Numerically accurate computational techniques for optimal estimator analyses of multi-parameter models

Lukas Berger[a]*, Konstantin Kleinheinz[a], Antonio Attili[a], Fabrizio Bisetti[b], Heinz Pitsch[a] and Michael E. Mueller[c]

[a]*Institute for Combustion Technology, RWTH Aachen University, 52056 Aachen, Germany;* [b]*Department of Aerospace Engineering and Engineering Mechanics, University of Texas at Austin, Austin, TX 78712-1085, USA;* [c]*Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, New Jersey, USA*

(*Received 4 August 2017; accepted 30 December 2017*)

Modelling unclosed terms in partial differential equations typically involves two steps: First, a set of known quantities needs to be specified as input parameters for a model, and second, a specific functional form needs to be defined to model the unclosed terms by the input parameters. Both steps involve a certain modelling error, with the former known as the irreducible error and the latter referred to as the functional error. Typically, only the total modelling error, which is the sum of functional and irreducible error, is assessed, but the concept of the optimal estimator enables the separate analysis of the total and the irreducible errors, yielding a systematic modelling error decomposition. In this work, attention is paid to the techniques themselves required for the practical computation of irreducible errors. Typically, histograms are used for optimal estimator analyses, but this technique is found to add a non-negligible spurious contribution to the irreducible error if models with multiple input parameters are assessed. Thus, the error decomposition of an optimal estimator analysis becomes inaccurate, and misleading conclusions concerning modelling errors may be drawn. In this work, numerically accurate techniques for optimal estimator analyses are identified and a suitable evaluation of irreducible errors is presented. Four different computational techniques are considered: a histogram technique, artificial neural networks, multivariate adaptive regression splines, and an additive model based on a kernel method. For multiple input parameter models, only artificial neural networks and multivariate adaptive regression splines are found to yield satisfactorily accurate results. Beyond a certain number of input parameters, the assessment of models in an optimal estimator analysis even becomes practically infeasible if histograms are used. The optimal estimator analysis in this paper is applied to modelling the filtered soot intermittency in large eddy simulations using a dataset of a direct numerical simulation of a non-premixed sooting turbulent flame.

Keywords: optimal estimator; multivariate adaptive regression splines; artificial neural networks; DNS; soot

1. Introduction

The concept of the optimal estimator enables a systematic decomposition of the total modelling error of any model into an irreducible and functional error [1]. When modelling, a set of known quantities always needs to be specified as input parameters for the model, which inevitably induces the irreducible error. In a second step, the functional error appears when a specific functional form for the unclosed terms as function of the input parameters is defined.

*Corresponding author. Email: l.berger@itv.rwth-aachen.de

For the quantification of these two errors, detailed data, either measured experimentally or generated via numerical simulation, are needed. For instance, turbulence models can be assessed by data from direct numerical simulations (DNS), where all turbulent length scales are properly resolved and no model is used. Based on such datasets, an empirical reference model with vanishing functional error is generated and compared to the actual model. These reference models are termed optimal estimators and simply use the mean of the unclosed term conditioned on certain input parameters as a model. When the same input parameters are used for the reference model and the model under investigation, the irreducible error of the actual model can be determined by the modelling error of the reference model.

In an optimal estimator analysis, the error decomposition enables the identification of the major source of modelling errors and a rigorous assessment of the potential for improvement of a certain model if either of the functional form or the input parameters are modified. If the total modelling error of a certain model is dominated by its irreducible error, a change of its functional form will not yield a significant improvement of the total modelling error. Furthermore, optimal input parameters for unclosed terms can be identified from a larger set of possible input parameters by a systematic analysis of irreducible errors.

The concept of optimal estimators was introduced by Moreau et al. [1], and has been applied in several studies [2–5] in order to assess and develop models for turbulent reacting flows. Typically, data from DNS are used for the analysis of such models since all turbulence and combustion length scales are resolved, allowing for a thorough analysis of the interactions of turbulence and combustion. However, DNS provide large amounts of data with a high level of detail so that data inference is only possible by means of systematic analysis tools, such as the concept of the optimal estimator [6]. It is worth noting here that optimal estimator analyses are data-driven and the results represent empirical findings. These empirical findings can then be used to guide the development of new physically motivated models.

In one study, Balarac et al. [2] analysed a DNS of forced isotropic turbulence in order to model the subfilter scalar dissipation rate, a quantity typically required in models for large eddy simulations (LES) for turbulent non-premixed combustion. The authors discuss three different models for the subfilter scalar dissipation rate: the local equilibrium assumption model (LEA), the large-scale strain rate tensor model (SRT), and the subfilter kinetic energy model (SKE). Each model uses a single, but different input parameter and the input parameter of the SKE model was found to yield the lowest irreducible errors. Thus, the SKE model reveals the largest potential for reducing the total modelling error if only the functional forms of the three models are modified, so the authors proposed a model formulation based on the input parameter of the SKE model whose total errors are lowest and close to the irreducible error. This work is a prototypical example of using the concept of optimal estimators to systematically reduce modelling errors.

In another study, Balarac et al. [3] compared a scale-similarity and dynamic Smagorinsky-type model for the subfilter scalar variance [3], another quantity typically required for LES of turbulent reacting flows. Smaller irreducible errors were obtained for the dynamic Smagorinsky-type model, but the scale-similarity model yielded lower total modelling errors. While one would conclude that the scale-similarity model might be superior, the optimal estimator analysis suggests that a potentially better model may be formulated when using the input parameters of the dynamic Smagorinsky-type model but changing its functional form. Conversely, the potential improvement of the scale-similarity model is limited since it already has relatively small functional errors. Such limitations in model improvement are difficult to analyse, but can be well disclosed in an optimal estimator analysis. Again, the optimal estimator analysis in this study has been limited to one input parameter. In both studies, Balarac et al. [2,3] used a histogram technique (HT) for the computation of the optimal estimators and irreducible errors. It will be shown that HTs accurately compute optimal estimators for one or two parameter models, but fail for larger numbers of input parameters.

In a third study, Vollant et al. [5] applied the optimal estimator approach to derive an empirical model for the subgrid-scale scalar flux from a DNS by using the optimal estimator of the subgrid-scale scalar flux based on seven input parameters directly as a model. Since the functional error of such models is zero, the total modelling error is only determined by its irreducible error. In an *a posteriori* analysis, they show that the surrogate model behaves very close to the filtered DNS results due to its relatively small modelling errors. However, the surrogate model fails if the flow conditions differ from those that are present in the data that are used to generate the surrogate model. As will be revealed in this work, the computation of optimal estimators with that many input parameters necessitates the usage of highly accurate computational techniques, so in their work, Vollant et al. used an artificial neural network (ANN) for the computation of the optimal estimator.

While the concept of optimal estimators has been used to systematically reduce modelling errors and identify the best sets of input parameters, little attention has been paid to the techniques themselves required for the practical computation of irreducible errors. Optimal estimator analyses involve the computation of conditional means and, particularly for large numbers of input parameters, the choice of technique for the computation of conditional means strongly affects the quantitative and possibly even the qualitative outcomes of the analysis. Therefore, different techniques for the numerical evaluation of optimal estimators are assessed in this work with a particular emphasis on a high-dimensional model-parameter space. A comparison among the following four techniques is presented: an HT [7], ANNs [8], multivariate adaptive regression splines (MARS) [9], and additive models based on a kernel method (AM) [10]. All these techniques belong to the class of non-parametric fitting methods, which are distinctly different from parametric fitting methods such as linear regression, where a certain pre-specified model is fit to the data. In the current study, we will show that the outcome of the optimal estimator analyses significantly varies if using different techniques.

In this paper, first, the concept of the optimal estimator is reviewed; then, the four non-parametric fitting methods for the practical computation of irreducible errors are discussed; finally, these methods are applied in an optimal estimator analysis of a turbulent reacting flow in order to demonstrate the impact of the non-parametric fitting technique on an optimal estimator analysis. A DNS dataset of a sooting turbulent flame [11,12] is used to study the filtered soot intermittency a priori in an optimal estimator analysis. The filtered soot intermittency is an indicator of how much subfilter volume is occupied by a soot volume fraction larger than a certain threshold, and hence characterises the subfilter spatial structure of the soot volume fraction. It is used in current subfilter models to model the evolution of soot quantities in LES [13]. Intermittency is chosen since it represents an excellent quantity for the investigation of multi-parameter models. The description of soot dynamics involves multiple soot moments such as the soot number density and the soot volume fraction. In an a priori analysis, further parameters that describe the subfilter spatial distribution of soot moments appear, so a large set of different parameters exists for a relevant application of the optimal estimator analysis.

In the employed DNS, the formation, growth, and transport of soot in a turbulent flame are analysed. Therefore, a detailed chemical mechanism, which includes the soot precursor naphthalene, is used and soot dynamics are described by the hybrid method of moments [14].

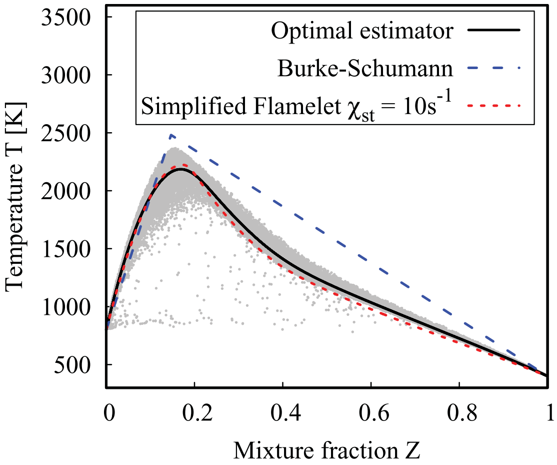


Figure 1. Dependency of temperature on mixture fraction for the employed DNS dataset (grey points) compared to the Burke–Schumann and a simplified flamelet model. (Colour online)

For the DNS configuration, a temporally evolving planar jet at a jet Reynolds number of 15,000 and atmospheric pressure is chosen. The central fuel jet contains *n*-heptane diluted with 85% (by volume) nitrogen at 400 K and is surrounded by a coflow of preheated air at 800 K. The stoichiometric mixture fraction is $Z_{st} = 0.147$. The domain is periodic in the streamwise and the spanwise directions while open boundaries are applied in crosswise direction. In this work, the DNS is studied at a time of about 9 jet times after its initialization, so soot-turbulence interactions have been able to develop sufficiently. For further details on the DNS database the reader is referred to the papers by Attili et al. [11,12].

2. The concept of the optimal estimator

The concept of the optimal estimator and its terminology are first demonstrated through an example of modelling the temperature T using the mixture fraction Z, for which the analysis will be performed with the DNS database described previously. Figure 1 shows the temperature conditioned on the mixture fraction for this particular DNS dataset and the respective conditional mean that has already been introduced as the optimal estimator. For the conditional mean, the temperature is conditioned on the mixture fraction and averaged over all of the DNS data. In addition, the comparison of the DNS data with two particular models is presented: the Burke–Schumann model [15] with an adiabatic flame temperature at stoichiometric mixture of 2481 K that has been computed from chemical equilibrium, and a steady flamelet solution [16] with a stoichiometric scalar dissipation rate of $\chi_{\mathrm{st}} = 10\,\mathrm{s}^{-1}$. Since only one flamelet is used in the latter, this model also depends only on the mixture fraction and although this is different from a steady flamelet model, where also the scalar dissipation rate appears as a parameter, this will be called a simplified flamelet model in the following. As expected, Figure 1 reveals a strong dependency of temperature on mixture fraction in the DNS. Furthermore, it is expected that the simplified flamelet model is more suitable for modelling the temperature in terms of the mixture fraction than the Burke–Schumann model. These findings are quantified in Table 1, where the average quadratic total modelling error over all N data points is shown for both models. This error

Table 1. Error decomposition for the simplified flamelet and Burke–Schumann models averaged over the DNS dataset.

Model	ϵ_{irr}^2	ϵ_{funct}^2	ϵ_{tot}^2
Burke–Schumann	$(41K)^2$	$(157K)^2$	$(162K)^2$
Simplified flamelet	$(41K)^2$	$(30K)^2$	$(51K)^2$
Optimal estimator	$(41K)^2$	$(0K)^2$	$(41K)^2$

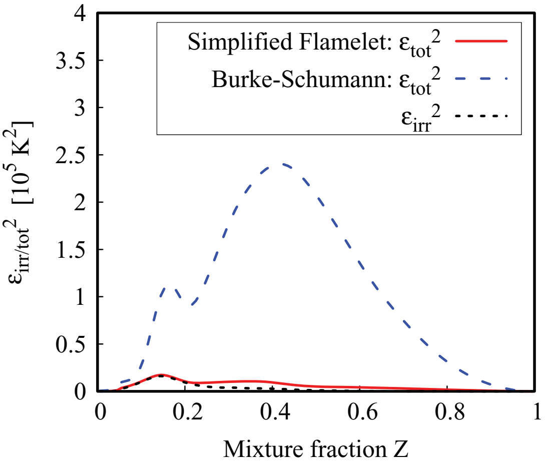


Figure 2. Error decomposition for the simplified flamelet and Burke–Schumann models for temperature plotted over mixture fraction for the employed DNS dataset. (Colour online)

is defined as:

$$\epsilon_{\rm tot}^2 = \frac{1}{N}\sum_{i=1}^{N}(T_{\rm DNS} - T_{\rm Model})^2. \tag{1}$$

In an optimal estimator analysis, the conditional mean in Figure 1 is termed the optimal estimator since by definition it represents the model with the lowest possible total modelling errors. No other model, based solely on mixture fraction, can approximate the temperature values of the DNS better. Therefore, the error associated with the optimal estimator is termed the irreducible error, and conceptually quantifies the degree of scatter with respect to the optimal estimator. Hence, this source of irreducible modelling error may only be changed if different or additional input parameters are used to describe the evolution of temperature in the temporally evolving jet, e.g. the local scalar dissipation rate.

Generally, any model that only uses mixture fraction as an input parameter to parametrise temperature has the same irreducible error. However, the total modelling errors may be different, as can be seen from Table 1 when comparing the Burke–Schumann and the simplified flamelet model. This is caused by the second source of modelling errors, which is induced by the particular functional form chosen to compute the temperature in terms of mixture fraction. The functional error is determined by the difference of the total and irreducible error. Thus, an optimal estimator analysis systematically decomposes the total modelling error into functional and irreducible error. In Figure 2, the error decomposition

Table 2. Irreducible errors for the temperature if using the axial velocity V, the mixture fraction Z, or the local scalar dissipation rate χ as input parameters.

Input parameters	ϵ^2_{irr}
none	$(415K)^2$
V	$(331K)^2$
Z	$(41K)^2$
Z, χ	$(33K)^2$

is shown with respect to mixture fraction for the Burke–Schumann and simplified flamelet model. It reveals that the total modelling error in the Burke–Schumann model is dominated by the specific functional form, while the simplified flamelet model is mostly affected by the irreducible error. Thus, the simplified flamelet model can only be significantly improved if more or different input parameters are used for the parametrisation of the temperature. According to Figure 2, the largest irreducible errors are observed around stoichiometry indicating an insufficient parametrisation of temperature in this region.

Table 2 shows different irreducible errors that are induced when parametrising the temperature by different input parameters, e.g. the mixture fraction, the local scalar dissipation rate, or the axial velocity. As expected, the irreducible error for the axial velocity is much larger compared to the irreducible error for the mixture fraction, which indicates that mixture fraction is a much more adequate parameter for the parametrisation of the temperature. Indeed, Table 2 shows that using the axial velocity for parametrisation is almost as bad as using a constant mean (case: none). However, Table 2 also quantifies that using the local scalar dissipation rate as additional input parameter to the mixture fraction only slightly improves parametrisation of the temperature. The DNS considered is far from extinction, so the effects of strain on temperature are relatively weak. Thus, an optimal estimator analysis also enables the rigorous assessment of different input parameters and can determine the optimal set of input parameters from a larger basis of input parameters.

Mathematically, the error decomposition of an optimal estimator analysis can be formally defined. Let Q be the quantity for which a particular model is required, e.g. temperature in the example above, and let Π be the set of input parameters chosen for the parametrisation of Q, e.g. mixture fraction in the example above, the corresponding conditional mean is:

$$g(\Pi) = \langle Q|\Pi\rangle, \tag{2}$$

where $g(\Pi)$ is termed the optimal estimator and $\langle\rangle$ represents an average over all of the DNS data. Note that Π can also represent a set of multiple input parameters. Let $M(\Pi)$ be a particular model formulation, e.g. the simplified flamelet model for the temperature in the example above, then the error decomposition described above becomes:

$$\begin{aligned}\epsilon^2_{tot}(\Pi) &= \langle[Q - M(\Pi)]^2|\Pi\rangle \\ &= \langle[Q - g(\Pi) + g(\Pi) - M(\Pi)]^2|\Pi\rangle\end{aligned}$$

$$
\begin{aligned}
&= \underbrace{\langle [Q - g(\Pi)]^2 | \Pi \rangle}_{\epsilon_{irr}^2} + \underbrace{[g(\Pi) - M(\Pi)]^2}_{\epsilon_{funct}^2} \\
&\quad + 2 \underbrace{[\langle Q|\Pi\rangle - g(\Pi)]}_{=0} \cdot [g(\Pi) - M(\Pi)] \\
&= \epsilon_{irr}^2(\Pi) + \epsilon_{funct}^2(\Pi). \qquad (3)
\end{aligned}
$$

In Equation (3), the error quantities ϵ_{tot}^2, ϵ_{irr}^2, and ϵ_{funct}^2 are determined as a function of the set of input parameters Π. For the previously discussed temperature models, the dependence of these error quantities on the mixture fraction is shown in Figure 2. However, if averaging these error values over the whole DNS dataset a single error value is obtained, which is the one shown in Table 1.

3. Computational techniques for optimal estimator analyses

The challenge in optimal estimator analyses is to accurately compute the conditional mean, such as $\langle Q|\Pi\rangle$, where the quantity Q, for which a model is required, is conditioned on the input parameter set Π. By definition, the conditional mean represents the local average of Q in parameter space Π. To determine such a local average, one has to use non-parametric fitting techniques, which are also referred to as smoothing techniques. In contrast to parametric fitting techniques such as linear regression, smoothing techniques do not require to fit a pre-specified model to the data so that a true local average can be obtained. Indeed, fitting a pre-specified model to the data is not even possible since the functional form of the optimal estimator is not known a priori.

First, the general idea of local averaging is presented, and second, the unavoidable inaccuracies in determining a local average for a given dataset are discussed. It will be shown that these unavoidable inaccuracies always lead to a certain error when determining the conditional mean by non-parametric fitting methods. These inaccuracies even increase if conditional means in a high-dimensional model-parameter space are computed and vary for different non-parametric fitting techniques. A third section will discuss the challenges of computing conditional means in a high-dimensional model-parameter space and a presentation of four different smoothing techniques, which will be subsequently referred to as ‘fitting techniques’, will be provided.

3.1. *Local averaging*

In practical implementations, local averaging of Q in parameter space Π always means averaging Q within a small non-zero volume in the vicinity of a given value of Π. If this volume is large, the average may not be regarded as local any more, and, if the volume is very small, the given dataset always becomes sparse so that no reliable local average can be determined. In both of these limits, the local average is not accurately determined so that there exists an optimal size for such an averaging volume. Figure 3 shows the behaviour of a fit with respect to the size of the averaging volume. Therefore, a dataset of two random variables Q and Π is generated that follows the relationship:

$$Q_i = g(\Pi_i) + \delta_i \qquad (4)$$

for each data point i where Π is uniformly distributed, δ follows a normal distribution with a zero mean, and the underlying functional relationship $g(\Pi)$ is shown in Figure 3 (black line). Thereby, a typical DNS scatter among two different quantities is approximated, but

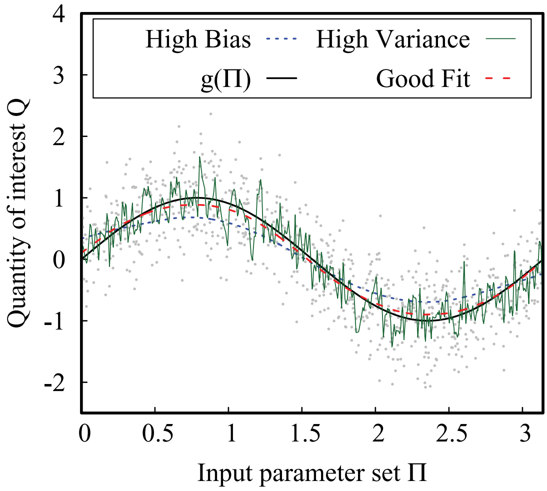


Figure 3. Variance-bias trade-off for a kernel smoother with different bandwidths h: $h_{\text{High Variance Fit}} = 0.01$, $h_{\text{High Bias Fit}} = 1.00$, $h_{\text{Good Fit}} = 0.50$. (Colour online)

in contrast to a DNS scatter, the optimal estimator is known a priori as $g(\Pi)$ in the limit of infinitely many data points. Figure 3 shows three different fits that try to reconstruct $g(\Pi)$ from the scattered data points whereby a fitting technique known as the kernel method is applied. It uses a constant averaging volume throughout the whole parameter space Π and the local average within the averaging volume is determined by a weighted mean. Details are provided in Section 3.3.2. Concerning the three different fits in Figure 3, one fit (blue dotted line) uses a relatively large averaging volume, yielding a very smooth fit, and in another case (green line), a relatively small averaging volume is chosen, resulting in a very unstable fit. For both cases, the underlying relationship from the data is not well reconstructed, as the smooth fit is very insensitive to the structures embedded in the data scatter, whereas the other fit is too sensitive and identifies statistically random patterns. As both fits significantly deviate from the conditional mean $g(\Pi)$, the errors associated with these fits will be larger than the irreducible error. An optimal fit (yellow dotted line) with a quadratic error close to the irreducible error would use a moderately large averaging volume such that the fit is sufficiently sensitive to actual variability in the data but also robust to random statistical variations. The fitting error resulting from the insensitivity of a technique is known as bias, and the fitting error that is induced by a too-sensitive technique is termed variance or overfitting. For any technique, both errors coexist, so all non-parametric fitting always involves a certain trade-off between bias and variance in order to keep the errors small that are related to the fitting technique itself [17].

3.2. *Bias and variance*

The existence of bias and variance for any technique affects the computation of optimal estimators and irreducible errors. Mathematically, the impact of bias and variance on the computation of irreducible errors is formally shown in the following. First, the concept of training and testing needs to be introduced: in order to properly assess the quality of a fit, it is necessary to split the dataset into two subsets. The fit is generated with one subset, which is known as training the fitting technique. Then, the second subset is used to compute the average quadratic error of the particular fit with respect to the data of the second subset, which is referred to as testing the fitting technique. This process ensures that errors due to overfitting are detected [17]. If one assessed the fits only on the basis of the training data, a fitting technique that is strongly overfitting would always yield the lowest errors as it identifies statistically random patterns in the data and approximates this particular dataset best. However, such an error measure would not be meaningful since, clearly, the 'high variance fit' does not capture the underlying relationship well according to Figure 3. In contrast, an error estimate based on the test dataset can detect overfitting since the underlying physics in the test dataset are the same, but statistically random patterns in the test dataset would be different [17].

As one typically only has limited data available, a compromise between the need to fit the data by a sufficiently large training set and the requirement to appropriately assess the fit is addressed by the cross-validation technique [7]. In this work, this technique is applied by splitting the data into two subsets and using one for training and one for testing. After having assessed the fit by the average quadratic error, the test dataset is used as a training dataset, and the second fit is assessed by the original training dataset. Thus, all data can be used for fitting and testing and the overall testing error is given as the average of both computed testing errors.

Let $E_{\{S\}}[...]$ be the operator that defines the averaging of the two testing errors $(\epsilon^2_{\text{Test}})^{S_1}$ and $(\epsilon^2_{\text{Test}})^{S_2}$ obtained for the first and second test dataset, S_1 and S_2, in the first and second stage of cross-validation, such that:

$$E_{\{S\}}\left[\left(\epsilon^2_{\text{Test}}\right)^S\right] = \frac{1}{2}\left[\left(\epsilon^2_{\text{Test}}\right)^{S_1} + \left(\epsilon^2_{\text{Test}}\right)^{S_2}\right] \quad \text{with} \quad S = S_1, S_2. \tag{5}$$

Let $g^{S_1}(\Pi)$ and $g^{S_2}(\Pi)$ be the fits based on the first and second training dataset during cross-validation and let $g(\Pi)$ be the optimal estimator, then, analogously to Equation (3), the error decomposition of the average testing error yields:

$$\begin{aligned}
\epsilon^2_{\text{Test}} &= E_{\{S\}}\left[\left(\epsilon^2_{\text{Test}}\right)^S\right] \\
&= E_{\{S\}}\left[\langle (g^S(\Pi) - Q)^2 | \Pi \rangle\right] \\
&= E_{\{S\}}\left[\langle (g^S(\Pi) - g(\Pi) + g(\Pi) - Q)^2 | \Pi \rangle\right] \\
&= \underbrace{E_{\{S\}}\left[(g^S(\Pi) - g(\Pi))^2\right]}_{\epsilon^2_{\text{Fit}}} + \underbrace{\langle (g(\Pi) - Q)^2 | \Pi \rangle}_{=\epsilon^2_{irr}}
\end{aligned} \tag{6}$$

The fitting error, which is induced if the fits $g^{S_1}(\Pi)$, $g^{S_2}(\Pi)$ do not coincide with the optimal estimator, can be further decomposed when defining an average fit $\hat{g}(\Pi) = E_{\{S\}}\left[g^S(\Pi)\right]$ among the two different training datasets S_1 and S_2.

$$\begin{aligned}
\epsilon^2_{\text{Fit}} &= E_{\{S\}}\left[(g^S(\Pi) - g(\Pi))^2\right] \\
&= E_{\{S\}}\left[(g^S(\Pi) - \hat{g}(\Pi) + \hat{g}(\Pi) - g(\Pi))^2\right]
\end{aligned}$$

$$
\begin{aligned}
&= E_{\{S\}}\left[(g^S(\Pi) - \hat{g}(\Pi))^2\right] \\
&\quad + \underbrace{E_{\{S\}}\left[2(g^S(\Pi) - \hat{g}(\Pi))(\hat{g}(\Pi) - g(\Pi))\right]}_{=0} \\
&\quad + (\hat{g}(\Pi) - g(\Pi))^2 \\
&= \underbrace{E_{\{S\}}\left[(g^S(\Pi) - \hat{g}(\Pi))^2\right]}_{\epsilon^2_{Variance}} + \underbrace{(\hat{g}(\Pi) - g(\Pi))^2}_{\epsilon^2_{Bias}} \qquad (7)
\end{aligned}
$$

Therefore, the fitting error can be decomposed into an error due to variance, which arises from the variability of the fit among the different training datasets, and an error due to bias. The bias results from a technique that does not allow for a sufficient flexibility, so throughout all training data a systematic error is induced. Evidently, fitting techniques are needed that only introduce small additional errors such that the testing error equals the irreducible error. However, practically, only the testing error can be computed and is taken as an estimate for the irreducible error. In Section 4, the impact of the fitting error on the computed irreducible error will be further discussed.

3.3. *Fitting techniques*

As demonstrated in Figure 3, fitting may be a challenging task that becomes even more challenging if a hypersurface in a high-dimensional space needs to be fitted to scattered data. In an optimal estimator analysis, such situations appear if multiple input parameters are used. One particular problem is that any dataset quickly becomes sparse in high dimensions, so conditional averages have to be evaluated on sparse data sets while keeping errors due to bias and variance small. For instance, a dataset of 10 billion data points conditioned on five input parameters will provide only about 100 points to evaluate the conditional mean in one parameter. This is referred to as the curse of dimensionality and is a fundamental issue that cannot be easily sidestepped [7]. One strategy to cope with the curse may be to locally reduce the dimensionality of the optimal estimator hypersurface by selecting only a small number of active variables out of Π in a particular region [7].

Additionally, it has been found [7,17,18] that the performance of the fitting techniques depends on the form of the optimal estimator itself, which is not known a priori. Banks et al. [18] compared the performance of different fitting techniques on the basis of different functional relationships that were embedded in the scattered data and found that some techniques are sometimes the best but sometimes also the worst. Therefore, as no method dominates all others over all possible datasets [17], a rigorous optimal estimator analysis requires at least two different fitting techniques in order to prove that the respective findings are not biased by the technique itself.

In this work, the optimal estimators and irreducible errors are computed with four techniques: a Histogram Technique (HT) [7], an Additive Model based on a kernel method (AM) [7], Multivariate Adaptive Regression Splines (MARS) [9], and Artificial Neural Networks (ANN) [8]. The HT is chosen as it is typically applied for the computation of conditional means. The AM simplifies high dimensional fitting with certain assumptions and is included in the analysis in order to demonstrate that high-dimensional fitting may not be addressed in such a simplified way. It will be shown that this method performs the worst for high-dimensional fits, but histograms also fail to accurately determine high-dimensional conditional means. Therefore, high-dimensional fits require the more sophisticated MARS and ANN techniques. Each of these four techniques is briefly described below.

3.3.1. *Histogram technique*

In HTs, the input parameter space Π is partitioned into pre-specified disjoint bins [7]. In this work, if Π contains more than one input parameter, the bins form a structured multi-dimensional grid such that the total number of bins is $(N_{\mathrm{Bins}})^{\dim(\Pi)}$, where N_{Bins} refers to the number of bins along one coordinate of the input parameter space Π and $\dim(\Pi)$ refers to the number of input parameters. The optimal estimator is then determined by the average of Q in each bin. The number of bins for each fit is found by an optimisation in which a small portion of the training data is held out in order to assess the fitting error based on the training data. This subset is referred to as a validation dataset. The fit with the lowest validation error is then assessed by the unused test dataset.

HTs provide an intuitive way of data fitting, but their ability of high-dimensional fitting is limited. First, local grid refinement for high-dimensional fits is challenging, so an equidistant grid determined by the number of bins is chosen in this work. In addition, histograms do not generalise well to high dimensions since any dataset becomes sparse in high dimensions, so test data may fall into bins where no training data are available. For these cases, either the mean of neighbouring bins needs to be used, or these data need to be simply neglected, which is the strategy applied in this work.

3.3.2. *Additive model and kernel methods*

The AM reduces the complexity of a high-dimensional fit to multiple one-dimensional fits by assuming additivity among the input parameters. In particular, AMs assume that the optimal estimator $g(\Pi)$ may be written as a sum of univariate functions that each depend only on one single input parameter π_k. Here, Π is written as the set of all input parameters π_k, such that $\Pi = \{\pi_1, \pi_2, ..., \pi_n\}$ with n input parameters. Thus, with the AM the optimal estimator is computed as:

$$g_{Fit}(\Pi) = \sum_{k=1}^{n} f_k(\pi_k), \tag{8}$$

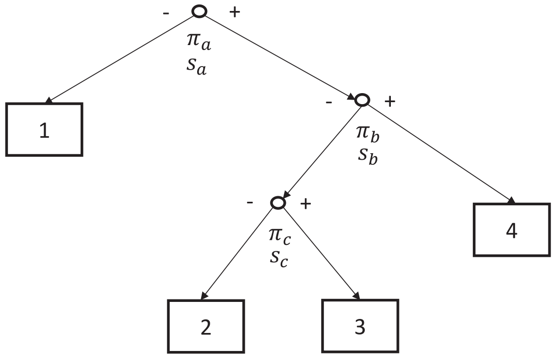
and the univariate functions $f_k(\pi_k)$ are determined by a one-dimensional fit, e.g. by a kernel method, which is used in this work.

Kernel methods determine the conditional mean $\langle Q|\Pi\rangle$ by a weighted average over all data points in the vicinity of a given value of Π. The radius within which the data points are considered for averaging is pre-specified and is termed the kernel bandwidth h. Let $K(\nu)$ be the weighting function for averaging the data points, let the index i mark the N data points of a given dataset, and for simplicity, the index k of $f_k(\pi_k)$ is dropped in the following, then the local average at the position $\hat{\pi}$ in parameter space is determined as:

$$f(\hat{\pi}) = \frac{\sum_{i=1}^{N} K(\pi_i - \hat{\pi}) Q_i}{\sum_{i=1}^{N} K(\pi_i - \hat{\pi})}. \tag{9}$$

In the literature, different weighting functions $K(\nu)$ exist, but typically, the Epanechnikov kernel is chosen [7]:

$$K(\nu) = \begin{cases} \frac{3}{4h}\left[1 - (\frac{\nu}{h})^2\right], & \text{if } |\nu| \leq h \\ 0, & \text{otherwise} \end{cases} \tag{10}$$



$$B_1 = f_-(x_{\pi_a} - s_a)$$
$$B_2 = f_+(x_{\pi_a} - s_a) \cdot f_-(x_{\pi_b} - s_b) \cdot f_-(x_{\pi_c} - s_c)$$
$$B_3 = f_+(x_{\pi_a} - s_a) \cdot f_-(x_{\pi_b} - s_b) \cdot f_+(x_{\pi_c} - s_c)$$
$$B_4 = f_+(x_{\pi_a} - s_a) \cdot f_+(x_{\pi_b} - s_b)$$

Figure 4. RPR tree with associated basis functions.

where h represents the kernel bandwidth. It may be chosen either by holding out a portion of the training data and picking the value of h for which the fit has the lowest quadratic error with respect to the held out data, or it can be specified a priori, which is known as a plug-in method for the kernel bandwidth [10]. In this work, h is pre-specified in order to minimise computational expense since fitting showed to be insensitive to h for a reasonably wide range of values.

AMs have the advantage that the complexity of high-dimensional fitting is reduced to multiple one-dimensional fits, which can be performed much more accurately. However, as a consequence, a strong assumption of additivity among the input parameters is made that does not necessarily apply to a given dataset. From all presented techniques, we will see that this technique performs the worst for high-dimensional fits since the assumptions made are not valid for the present optimal estimator analysis.

3.3.3. *Multivariate adaptive regression splines*

MARS have been introduced by Friedman [9] and are closely related to the recursive partitioning regression (RPR) [7] technique. The latter can intuitively be understood by its geometrical interpretation. In a first step, the whole parameter space domain is split into two daughter regions, and, in each region, a constant function is fit to the data. In a next step, one region is split again into two daughter regions, and constant functions are fit to the daughter regions and so forth. For each step, an optimisation with respect to the placement of the split is performed with the objective of minimising the mean error between the training data and the region-wise constant functions of RPR [9]. The domain splitting may be understood as growing a tree, shown in Figure 4. In this case, three variables are chosen as input parameters: $\pi_{a,b,c}$ indicates the specific split variable and $s_{a,b,c}$ represents the variable's value at which the domain is split in each step. Typically, in RPR algorithms, first large trees are generated, and afterwards they are pruned back in order to avoid overfitting [7].

Therefore, RPR efficiently exploits low dimensionality because variables that have little impact are less likely to be picked for splitting [9].

In MARS, the concept of regions and splitting is replaced in a mathematical sense by addition and multiplication. Each subdomain m is associated with a function B_m that is zero outside and non-zero inside the subdomain, yielding the following approximation of $g(\Pi)$:

$$g_{Fit}(\Pi) = \sum_{m=1}^{M} a_m B_m(\Pi), \tag{11}$$

where M indicates the number of all domains and a_m is a fitting parameter.

In particular, when splitting a domain, the existing domain function, which is termed the parent function, is multiplied by a left and a right sided basis function (termed f_- and f_+ in Figure 4), yielding two new daughter domain functions. For instance, the right sided basis function is given by:

$$f_+(x-s) = \begin{cases} 0 & x \le s_- \\ c_1(x-s_-)^2 + c_2(x-s_-)^3 & s_- \le x \le s_+ \\ x-s & x \ge s_+ \end{cases} \tag{12}$$

where the variables $s_\pm$ are chosen close to the splitting point s so that the basis function is zero for almost all the domain left of the splitting point and an almost linear function right of the splitting point. Thus, a multiplication of a given domain function with the right sided basis function generates a daughter domain that comprises the parent domain to the right of the splitting point. The cubic transitioning term around the splitting point and its respective constants c_1, c_2 are chosen such that the first derivative is continuous at s_- and s_+.

At the beginning of the fitting procedure, MARS assumes a function basis $\{B_m\}$ that only contains a constant domain function $B_0 = 1$. This function is split into two daughter functions by the multiplication of B_0 with the left and right sided basis functions, yielding the new function basis $\{B_0, B_1, B_2\}$. This basis is extended by further splitting procedures while in contrast to RPR, the parent function is always kept in the actual function basis $\{B_m\}$. Thus, the actual function basis $\{B_m\}$ may contain certain basis functions that overlap regions of others. At each step, an optimisation with respect to the placement of the split is performed with the objective of minimising the mean error between the training data and the fit that would be generated by the respective split placement. Note that a basis function B_m cannot be split twice along a coordinate π_k. However, B_0 is always available for splitting along any coordinate.

As each domain function fits an almost linear function to the data, the successive domain splitting leads to an almost continuous piecewise linear fit. Thereby, the splitting points and the slope of the piecewise lines a_m represent the respective fitting constants. Analogously to RPR, the extension of a current basis $\{B_m\}$ in MARS can be understood as growing a tree that is pruned back after sufficient splitting to avoid overfitting. More details may be found in the original paper by Friedman [9].

For this study the MARS implementation of Friedman (Version 3.5 in Fortran 77 [9,19–21]) is used.

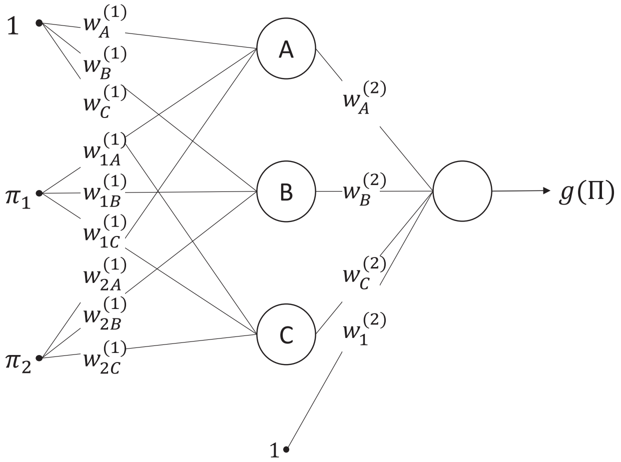


Figure 5. Single hidden layer Artificial Neural Network.

Throughout this study, MARS is found to accurately predict optimal estimators even for high-dimensional fits, but coming along with a significantly increased computational expense particularly for high-dimensional fits.

3.3.4. *Artificial neural networks*

ANNs were originally developed to better understand the activity of physiological neurons. Mathematically, ANN are composed of nodes that are often organised in layers, and every single node is interconnected to all nodes in the preceding and the following layer. Generally, many different network architectures are possible, and a single hidden layer neural network is shown in Figure 5. For such a network, the training data are inserted into the network, then a weighted sum of these inputs according to the weights $\{w_k^{(1)}\}$ is transferred to the nodes of the hidden layer, processed within the node by an activation function ϕ, and finally, these outputs are transferred by the weights $\{w_k^{(2)}\}$ to the output node, which represents the optimal estimator $g(\Pi)$ [8]. The hidden layer is termed hidden since only the linear combination of its outputs weighted by $\{w_k^{(2)}\}$ is actually seen. The network architecture and number of neurons is predefined, and the data are used for training the weights $\{w_k^{(1)}\}$ and $\{w_k^{(2)}\}$.

Mathematically, fitting by a neural network may be seen as a parametric fit since the optimal estimator is approximated by:

$$g(\Pi) = \sum_{k=1}^{N_{\text{Neurons}}} w_k^{(2)} \phi \left(\sum_{j=1}^{\dim(\Pi)} w_{jk}^{(1)} \pi_j + w_k^{(1)} \right) + w_1^{(2)}, \tag{13}$$

where N_{Neurons} represents the number of neurons in the hidden layer and dim(Π) refers to the number of input parameters in Π. However, neural networks are much less restrictive than typical parametric fits as they contain a rich enough class of functions and allow for a sufficient flexibility in data fitting [7].

For this investigation, the neural network toolbox of MATLAB with a single hidden layer neural network, a hyperbolic tangent sigmoid activation function,

$$\phi(\xi) = \frac{2}{1 + \exp(-2\xi)} - 1, \tag{14}$$

and a Bayesian regularisation training is used. The initial weights of the neural network $\{w_k^{(1,2)}\}$ are randomly set, which, if repeating the same optimal estimator analysis, did only cause deviations of the irreducible errors smaller than 0.1%. For an optimal choice of the number of neurons, a small portion of the training data is held out and is used for the assessment of the fit. After finding the optimal number of neurons with the smallest training error, the final fit is assessed by the unused test dataset.

Throughout this study, ANN are found to accurately predict optimal estimators even for high-dimensional fits, and additionally showed to be computationally efficient since neural networks can partially evade the curse of dimensionality [7].

4. Results and discussion

In this section, the impact of a particular fitting technique on determining conditional means and irreducible errors in an optimal estimator analysis is investigated.

For the analysis here, the quantity of interest is the filtered soot intermittency, which is studied in the DNS database described above for a sooting turbulent non-premixed jet flame. The soot intermittency is defined as the probability of finding a soot volume fraction that is less than a certain threshold. Following Qamar et al. [22], this threshold is given by the detection limit of experimental devices and is found to be 0.1 ppb. The filtered soot intermittency then represents the fraction of the subfilter volume that has a soot volume fraction below the respective threshold or, physically, the subfilter structure of the soot volume fraction.

Mathematically, the unfiltered intermittency of the unfiltered soot volume fraction f_V is represented by a Heaviside function:

$$I(f_V) = H(f_V^* - f_V) \quad f_V^* = 0.1\text{ppb}. \tag{15}$$

Concerning LES, the filtered intermittency can be expressed by convoluting the intermittency with the soot subfilter probability density function (PDF) $\mathcal{P}(f_V)$ [13] that describes the subfilter distribution of the soot volume fraction:

$$\overline{I} = \int \mathcal{P}(f_V)\, H(f_V^* - f_V) df_V. \tag{16}$$

This subfilter PDF can also be used to define both the filtered soot volume fraction and its higher-order subfilter moments:

$$\overline{f_V} = \int f_V \mathcal{P}(f_V) df_V \tag{17}$$

$$\phi_k = \int (f_V - \overline{f_V})^k \mathcal{P}(f_V) df_V. \tag{18}$$

Vice versa, the subfilter PDF is parametrised by its subfilter moments:

$$\mathcal{P}(f_V) = \mathcal{P}(f_V|\overline{f_V}, \phi_2, \phi_3, ...). \tag{19}$$

Here, the objective of the optimal estimator analysis is to determine which set of these subfilter moments best describes the filtered intermittency or, equivalently, the subfilter PDF:

$$\overline{I}(\overline{f_V}, \phi_2, \phi_3, ...) = \int \mathcal{P}(f_V|\overline{f_V}, \phi_2, \phi_3, ...)\, H(f_V^* - f_V) df_V. \tag{20}$$

4.1. *Technical details*

Before discussing the impact of a particular fitting technique on determining conditional means and irreducible errors, all additional technical details for the subsequent analyses are presented.

If not mentioned differently, the following is applied to all of the subsequent optimal estimator analyses. All optimal estimators are computed by ANNs and the irreducible errors are averaged with respect to the whole DNS dataset according to Equation (1). In order to guarantee a proper non-parametric fit, the dataset needs to be preprocessed before fitting, so optimal estimators are always determined with respect to the logarithm of the input parameters instead of using the input parameters directly without preconditioning them. This will be discussed in Section 4.3. For the subsequent optimal estimator analysis not all data are considered, since non-parametric fitting using the aforementioned techniques proved computationally infeasible if all data points are used. In total, the DNS dataset comprises about 500 million data points, but, for the following analyses, only about 400,000 data points are used. Therefore, first the DNS data are filtered by a box filter with a filter size Δ of $\Delta = 43\eta$, where $\eta = 110\,\mu$m is the Kolmogorov length of the turbulent flame [11,12]. Filtering is done for each single data point, so the filtering volumes overlap and each filtering volume comprises 51^3 data points. From the filtered data, only every 11th data point in each spatial direction is considered for the optimal estimator analyses, so the dataset is reduced by a factor of 11^3. In Figure 6, the irreducible errors for the filtered intermittency conditioned on eight different input parameter sets are shown for three differently large datasets. The three datasets comprise 0.4 million, 1.3 million, and 17 million data points, and are generated by only considering every 11th, seventh or third data point of the filtered field in each spatial direction. The different input parameter sets are described in Table 3 and consist of the first eight subfilter moments of the soot volume fraction ϕ_k. The first input parameter set Π_1 only contains the filtered soot volume fraction $\overline{f_V}$, and each subsequent parameter set includes the next higher subfilter moment of the soot volume fraction, e.g. Π_2 contains the filtered soot volume fraction $\overline{f_V}$ and its subfilter variance ϕ_2. However, the dataset reduction does not affect the computation of the irreducible errors according to Figure 6 even for cases that require the computation of high-dimensional conditional means. Note that the computation of the missing values in Figure 6 was computationally infeasible, i.e. irreducible errors for two and more parameters could not be computed if 17 million data points are used. Additional filter sizes of $\Delta = 26\eta$, $\Delta = 9\eta$, and $\Delta = 2.5\eta$ have been investigated, but results are found to not significantly vary with respect to the filter size.

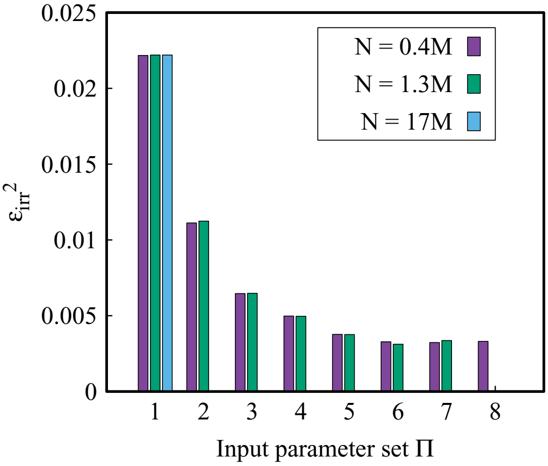


Figure 6. Irreducible errors for the filtered soot intermittency computed with ANN for three different subsets of the full DNS dataset, each containing N data points. The input parameters $\Pi_{1,..,8}$ are given in Table 3. Determination of missing values was computational infeasible. (Colour online)

4.2. *Optimal estimator analysis*

In this optimal estimator analysis, the filtered intermittency is conditioned on eight different input parameter sets $\Pi_{j=1..8}$. The input parameter sets are listed in Table 3, which has been described in the previous section, and the irreducible errors for these input parameter sets are shown in Figure 6. As expected, additional input parameters lead to reduced irreducible errors as more information about the subfilter distribution is used for the parametrisation of the subfilter PDF. However, the addition of increasingly high order subfilter moments does not significantly improve the parametrisation of the subfilter PDF as the irreducible errors level off for Π_6, Π_7, and Π_8. This behaviour of the irreducible errors will be further discussed in Section 4.5.

In the following subsections, we will show that this analysis is affected by two aspects related to the practical computation of conditional means. First, the findings of this optimal estimator analysis change if different fitting techniques are employed, and second, the manner in which the data are preprocessed before being used for fitting is important.

Table 3. Input parameter sets for optimal estimator analysis in Figures 6, 7, and 10.

Π_i	Parameter set Π_i
1	$\{\overline{f_V}\}$
2	$\{\overline{f_V}, \phi_2\}$
3	$\{\overline{f_V}, \phi_2, \phi_3\}$
4	$\{\overline{f_V}, \phi_2, ..., \phi_4\}$
5	$\{\overline{f_V}, \phi_2, ..., \phi_5\}$
6	$\{\overline{f_V}, \phi_2, ..., \phi_6\}$
7	$\{\overline{f_V}, \phi_2, ..., \phi_7\}$
8	$\{\overline{f_V}, \phi_2, ..., \phi_8\}$

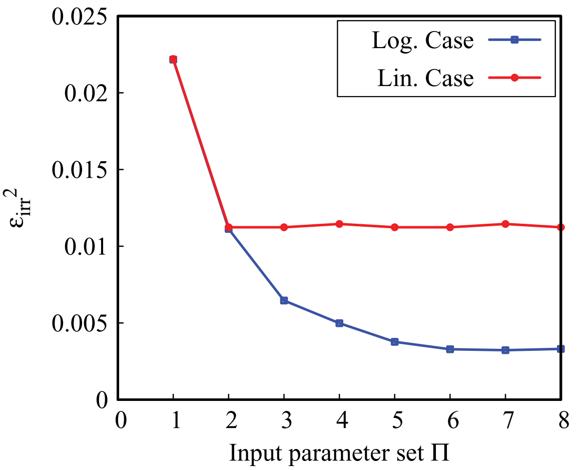


Figure 7. Irreducible error for the filtered soot intermittency with two different data preprocessing techniques computed with ANN. The input parameters $\Pi_{1,..,8}$ are given in Table 3. (Colour online)

4.3. *Effect of data preprocessing*

If and how the input parameters are preconditioned before fitting should not matter. However, Figure 7 shows the irreducible errors of the filtered intermittency conditioned on the different subfilter moments for two cases: the raw data are directly inserted into the fitting procedure (linear case), and the logarithm of the raw data is inserted into the fitting procedure (logarithmic case). It is worth noting that all subfilter moments ϕ_k of the soot volume fraction appeared to be positive, so the logarithm of any subfilter moment ϕ_k is well defined. For one- and two-dimensional fits, the irreducible errors are the same irrespective of the preprocessing procedure, but, for all higher-dimensional fits, the irreducible errors of the linear case remain constant, while in the logarithmic case irreducible errors keep decreasing. As the value of the conditional mean of the filtered intermittency is supposed to remain unaffected for a given position in Π-space, the irreducible error should also remain unaffected by the preconditioning. Recalling Equation (6), the computed irreducible errors in the linear case must then be strongly affected by errors arising from the fitting technique itself.

For a deeper understanding of this observation, Figure 8 shows the filtered intermittency conditioned on the filtered soot volume fraction and Figure 9 shows the filtered intermittency conditioned on the logarithm of the filtered soot volume fraction. Additionally, the PDF of the DNS data, distinct from the subfilter PDF of the soot volume fraction, with respect to the filtered soot volume fraction and the logarithm of the filtered soot volume fraction is shown. The filtered soot volume fraction covers many orders of magnitude, and a significant portion of the data possesses relatively low filtered soot volume fractions. Therefore, in the linear case a strong peak of the PDF is observed close to the origin which makes it very challenging for the fitting technique to resolve the embedded structures in these regions and results in a very large bias error. In contrast, in the logarithmic case, the data distribution is more uniform, and the resulting bias error is significantly reduced.

Summarising, an adequate preprocessing of the data significantly reduces fitting errors since it may lead to a more uniform distribution of the data points in parameter space, but the manner in which the data need to be preprocessed remains specific to each dataset.

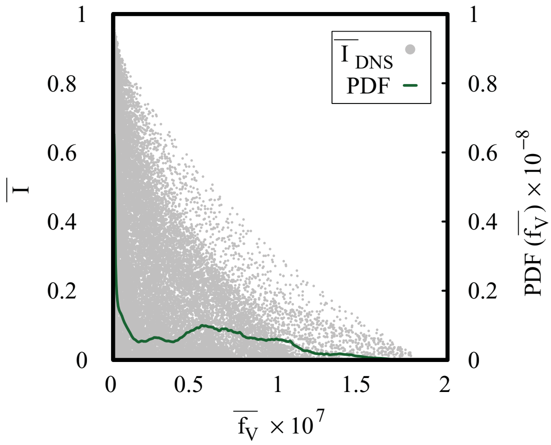


Figure 8. Effects of data preprocessing on the distribution of the DNS data: linear case. (Colour online)

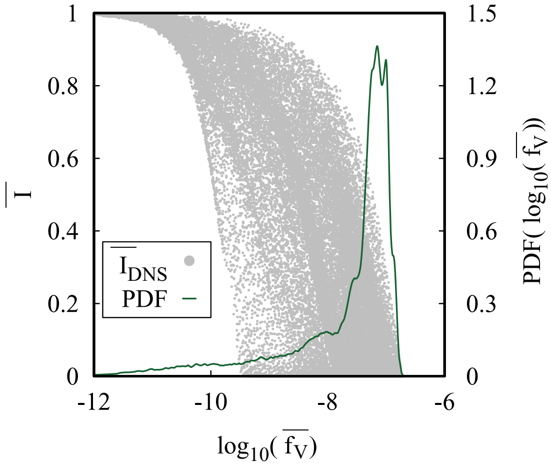


Figure 9. Effects of data preprocessing on the distribution of the DNS data: logarithmic case. (Colour online)

4.4. *Effect of fitting technique*

Figure 10 shows the same optimal estimator analysis as presented in Figure 6, but carried out with each of the four different introduced techniques: HT, AM, MARS, and ANN. The fitting techniques result in very different irreducible errors, with the differences increasing with the dimensionality of Π. Recalling Equation (6), these deviations must be due to the particular fitting technique since the irreducible error is the same for a given parameter set Π. ANN has the lowest errors related to the fitting technique while AM generates the largest errors for high-dimensional fits. These results simply reflect the curse of dimensionality, with which some techniques can cope better than others. However, all techniques yield identical results for one-dimensional fits, which verifies the implementation of the different fitting techniques, and indicates that the fitting error is negligibly small compared to the irreducible error for a single input parameter.

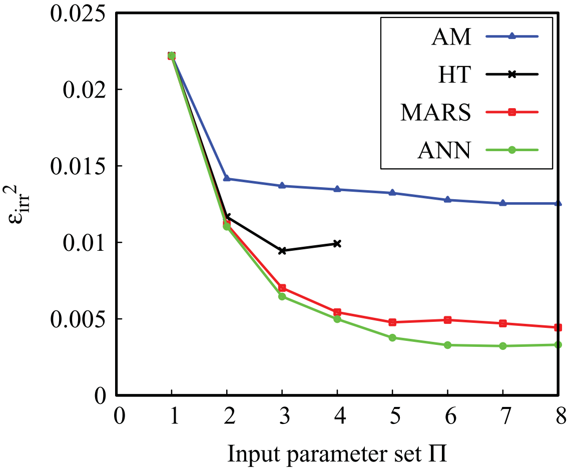


Figure 10. Irreducible errors for the filtered soot intermittency computed by different fitting techniques. The input parameters $\Pi_{1,..,8}$ are given in Table 3. (Colour online)

4.4.1. *Histogram technique*

Figure 10 shows reasonable agreement of the results computed with HT compared with those of MARS and ANN up to two parameters even though very small deviations of HT are seen for two parameters with respect to MARS and ANN. For three and four parameters, the HT introduces significant fitting errors and deviates significantly from MARS and ANN. As previously mentioned, these fitting errors result from the poor capability of HT to generalise to high-dimensional problems, since any dataset becomes sparse in high dimensions, so that test data may fall into bins where no training data are available. Furthermore, grid refinement in high dimensions is challenging for HT, so the adaptivity of HT to the high-dimensional data structure is limited. Moreover, no irreducible errors have been computed for dimensions higher than four, since HTs require the computation of a local average in $(N_{\mathrm{Bins}})^{\dim(\Pi)}$ bins, which quickly reaches memory limits.

4.4.2. *Additive model*

For the AM, the irreducible error in one dimension is well captured, but significant errors are introduced for higher dimensional fits. This results from the additivity assumption in Equation (8) that the optimal estimator hypersurface $g(\Pi)$ may be decomposed into sums of univariate functions of the single input parameters. By definition, the additivity assumption neglects any interactions between the parameters, so the fitting errors increase rapidly as the dimensionality of the input parameter set is increased. Unless the input parameters are known to have independent effects a priori, AM should not be used for high-dimensional input parameter sets.

4.4.3. *Multivariate adaptive regression splines*

In this work, the splitting tree of MARS is allowed to contain at most $M_{max} = 500$ splits. This maximum number of splits M_{max} was chosen as it was sufficiently larger than the number of functions in the final basis $\{B_m\}$ after the splitting tree is pruned back. Compared to ANN, slightly higher quadratic errors are computed by MARS for the input parameter sets $\Pi_{i>3}$. This and particularly the increase of irreducible errors from Π_5 to Π_6 relates to deficiencies of MARS as a constant or decreasing irreducible errors is expected. It is difficult to say where these discrepancies arise from, but considering that high-dimensional fitting is a challenging task in itself and that the deviations between MARS and ANN compared to the other techniques are relatively small, MARS may still be regarded as an appropriate tool for optimal estimator analyses of low and moderately high-dimensional input parameter sets.

4.4.4. *Artificial neural network*

For the fits by ANN, a single hidden layer network is chosen. Compared to MARS, ANN was significantly more computationally efficient. Throughout the whole analysis, the irreducible errors computed by ANN always yield the lowest values, so according to Equation (6) errors induced by the technique itself are also the smallest, even though such errors may potentially still dominate the computed quadratic errors. However, since the deviations between MARS and ANN compared to the other techniques are relatively small, one may assume that the errors shown in Figure 10 represent the irreducible error as both techniques have very different fitting concepts, but still yield almost identical results.

4.4.5. *Remarks*

Low-dimensional fits are accurately predicted by HT, MARS, and ANN but high-dimensional fits are only reasonably accurately predicted by MARS and ANN, as HT suffers severely from the curse of dimensionality. Moreover, the computation of high-dimensional fits by HT was not feasible due to the enormous memory requirements of HT. The computations of MARS and ANN remains feasible even for high dimensional fits, but ANN is more computationally efficient. AM only yields reasonable results for one-dimensional fits. These findings suggest that an optimal estimator analysis should always be performed with at least two fitting methods, specifically ANN and MARS, to prove that the results are technique-independent.

4.5. ***Non-vanishing irreducible errors***

As a final point, in Figure 6, the irreducible errors do not appear to decay to zero as the dimensionality of the input parameter set increases. Intuitively, one could expect the subfilter PDF to be perfectly parametrised if sufficiently many integer subfilter moments are used for parametrisation. However, Figure 6 reveals that adding high-order subfilter moments to the set of input parameters does not significantly improve the parametrisation of the subfilter PDF, and the irreducible errors even saturate. In Figure 11, the subfilter PDFs for five different subfilter volumes are shown. All five volumes have the same values of the first eight subfilter moments ($=\Pi_8$); however, the subfilter PDFs have different shapes. The PDFs have been computed by means of a kernel method such that

$$\mathcal{P}(\hat{f}_V) = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} K\left(\log_{10}(f_V|_i) - \log_{10}(\hat{f}_V)\right). \tag{21}$$

determines the subfilter PDF for a given value of the soot volume fraction $\hat{f}_V$.$f_V|_i$ represents the unfiltered soot volume fraction of the N_{SV} data points within a subfilter volume and $K(\nu)$ is the kernel function from Equation (10).

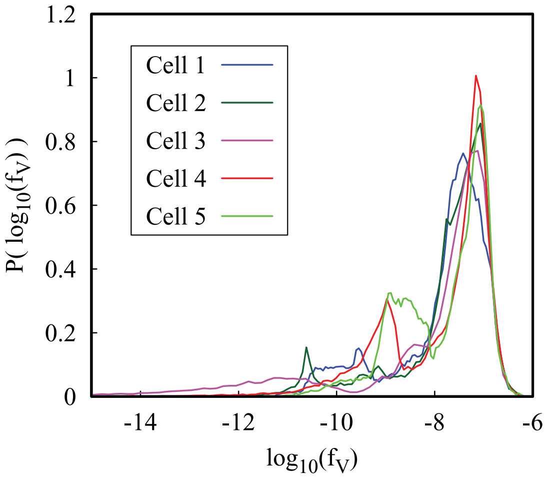


Figure 11. Five subfilter PDFs evaluated from the employed DNS with the same first eight, low-order subfilter moments. (Colour online)

As shown in Figure 11, the soot volume fraction is distributed over many orders of magnitude. Consequently, the subfilter moments are biased towards values of large soot volume fractions within the subfilter volume, and increasing the order simply increases this bias. Therefore, integer-order subfilter moments are insufficient for parametrising the subfilter PDF for small values of the soot volume fraction, as can be seen from Figure 11. A different choice of input parameters may result in an improved parametrisation of the subfilter PDF in these regions. For instance, one can define fractional moments of the subfilter PDF as:

$$\xi_k = \int (f_V - \overline{f_V})^{1/k} \mathcal{P}(f_V) df_V \quad k \in \mathbb{N}. \tag{22}$$

These moments place a stronger weight on small values of the soot volume fraction. In Figure 12, irreducible errors are compared when conditioning the filtered intermittency on different input parameter sets, where each subsequent parameter set includes either the next higher integer or fractional subfilter moment of the soot volume fraction starting from $\Pi_1 = \overline{f_V}$. The input parameter sets are described in Table 4. Figure 12 shows that irreducible errors can be significantly reduced using fractional moments compared to the previously defined integer-order subfilter moments ϕ_k.

Note that, mathematically, a necessary condition for a complete parametrization of a distribution by all of its integer moments is a positive radius of convergence of the moment generating function [23]. As the subfilter PDFs appear to be approximately log-normally distributed in Figure 11, this requirement is not fulfilled, which causes the non-vanishing irreducible error in Figure 6.

However, even though fractional moments are found to be more adequate for the parametrisation of the subfilter PDF, it might be difficult to use them for modelling since the derivation of a transport equation for a fractional subfilter moment is problematic.

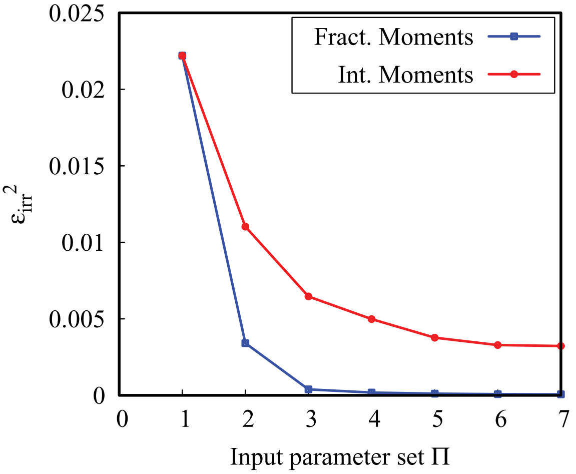


Figure 12. Irreducible errors for the filtered soot intermittency using integer-order and fractional-order subfilter moments as input parameters computed by ANN. The input parameters $\Pi_{1,\dots,8}$ are given in Table 4. (Colour online)

Starting from the transport equation of the unfiltered soot volume fraction [13]:

$$\frac{\partial f_V}{\partial t} + \frac{\partial (u_j^* f_V)}{\partial x_j} = \dot{M}, \tag{23}$$

where u_j^* represents the flow velocity corrected by the effect of thermodiffusion, and $\dot{M}$ represents a source term, the transport equation for a fractional soot volume fraction with exponent $k \in (0, 1)$ yields:

$$\frac{\partial f_V^k}{\partial t} + \frac{\partial (u_j^* f_V^k)}{\partial x_j} + (k-1) f_V^k \frac{\partial u_j^*}{\partial x_j} = k f_V^{k-1} \dot{M}. \tag{24}$$

Filtering Equation (24) results in the transport equation of a fractional subfilter moment ξ_k. However, Equation (24) is only well defined if $f_V \neq 0$, so the transport equation of the

Table 4. Input parameter sets for Figure 12. If fractional moments are used as input parameters $A_k = \xi_k$. If integer moments are used as input parameters $A_k = \phi_k$.

Π_i	Parameter set Π_i
1	$\{\overline{f_V}\}$
2	$\{\overline{f_V}, A_2\}$
3	$\{\overline{f_V}, A_2, A_3\}$
4	$\{\overline{f_V}, A_2, \dots, A_4\}$
5	$\{\overline{f_V}, A_2, \dots, A_5\}$
6	$\{\overline{f_V}, A_2, \dots, A_6\}$
7	$\{\overline{f_V}, A_2, \dots, A_7\}$

fractional subfilter moment ξ_k is only well defined for regions where $f_V \neq 0$ throughout the whole subfilter volume.

5. Conclusion

In this work, the non-negligible impact of the computational techniques on an optimal estimator analysis has been demonstrated. In an optimal estimator analysis, accurate non-parametric fits of scattered data are crucial since they are needed for the computation of optimal estimators and irreducible errors. Therefore, computational techniques are required that are capable of accurately computing conditional means over high-dimensional input parameter sets. Mathematically, it has been rigorously shown that the computed irreducible error may deviate from the data intrinsic irreducible error by an error induced by the computational technique itself.

Four different computational techniques have been assessed in an optimal estimator analysis of the filtered soot intermittency: HT, an AM which uses a kernel method, MARS, and ANN. Large deviations of the computed irreducible errors are found among the different techniques that increase when computing irreducible errors for an increased number of input parameters.

The HT showed a reasonable performance for one- and two-dimensional fits but did not generalise well to high dimensions. Moreover, the computation of high-dimensional fits by histograms was not feasible due to the enormous memory requirements of the HT. AMs only showed satisfactory results for one-dimensional fits. MARS and ANN performed well for any case, although MARS proved to be slightly less accurate than ANN if more than five input parameters were used. In addition, ANN was found to be computationally more efficient than MARS. Finally, it is also shown that appropriately preprocessing the data before fitting, e.g. by using the logarithm of the original data, significantly improves the results.

Optimal estimator analyses may be applied to the analysis of any quantity of interest as long as a sufficient and adequate dataset exists. However, it is of particular interest for LES model development from DNS data, as such datasets specifically require systematic analysis tools. The computational effort and the magnitude of the discrepancies among the different techniques may vary for optimal estimator analyses of other quantities compared to the present findings, but the conclusions of the present study remain unaffected. Optimal estimator analyses should always be carried out by at least two different computational techniques, specifically ANN and MARS, to prove that the results are not biased by the technique itself.

Acknowledgements

The analyses presented in this article were performed in part on computational resources supported by the Princeton Institute for Computational Science and Engineering (PICSciE) and the Office of Information Technology’s Research Computing department at Princeton University and in part by the Gauss Centre for Supercomputing e.V. on the GCS Supercomputer SuperMuc at Leibniz Supercomputing Centre in Munich. We acknowledge the KAUST Supercomputing Laboratory (KSL) for providing the computational time on the supercomputer “Shaheen” used to perform the DNS.

Funding

Generous support of the Deutsche Forschungsgemeinschaft (DFG) and the Research Association for Combustion Engines (FVV) under grant numbers PI 368/6-2 (DFG) and 6012390 (FVV) for L. B., K. K., and H. P. is gratefully acknowledged. A. A. and H. P. gratefully acknowledge generous support of the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 695747).

References

[1] A. Moreau, O. Teytaud, J. P. Bertoglio, *Optimal estimation for large-eddy simulation of turbulence and application to the analysis of subgrid models*, Phys. Fluids 18, 105101 (2006).

[2] G. Balarac, H. Pitsch, V. Raman, *Modeling of the subfilter scalar dissipation rate using the concept of optimal estimators*, Phys. Fluids 20, 091701 (2008).

[3] G. Balarac, H. Pitsch, V. Raman, *Development of a dynamic model for the subfilter scalar variance using the concept of optimal estimators*, Phys. Fluids 20, 035114 (2008).

[4] Y. Fabre, G. Balarac, *Development of a new dynamic procedure for the Clark model of the subgrid-scale scalar flux using the concept of optimal estimator*, Phys. Fluids 23, 115103 (2011).

[5] A. Vollant, G. Balarac, C. Corre, *Subgrid-scale scalar flux modelling based on optimal estimation theory and machine-learning procedures*, J. Turb. 18 (2017), pp. 854–878.

[6] P. Trisjono, H. Pitsch, *Systematic analysis strategies for the development of combustion models from DNS: A review*, Flow Turb. Combust. (2015), pp. 231–259.

[7] B. Clarke, E. Fokoué, H.H. Zhang, *Principles and Theory for Data Mining and Machine Learning*, Springer, New York , 2009.

[8] C. M. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, 1995.

[9] J. H. Friedman, *Multivariate adaptive regression splines*, Ann. Stat. 19 (1991), pp. 1–141.

[10] W. Härdle, *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, 1990.

[11] A. Attili, F. Bisetti, M. E. Mueller, H. Pitsch, *Formation, growth and transport of soot in a three-dimensional turbulent non-premixed jet flame*, Combust. Flame 161 (2014), pp. 1849–1865.

[12] A. Attili, F. Bisetti, M. E. Mueller, H. Pitsch, *Effects of non-unity Lewis number of gas-phase species in turbulent nonpremixed sooting flames*, Combust. Flame 166 (2016), pp. 192–202.

[13] M. E. Mueller, H. Pitsch, *Large eddy simulation subfilter modeling of soot-turbulence interactions*, Phys. Fluids 23, 115104 (2011).

[14] M. E. Mueller, G. Blanquart, H. Pitsch, *Hybrid method of moments for modeling soot formation and growth*, Combust. Flame 156 (2009), pp. 1143–1155.

[15] S. P. Burke, T. E. W. Schumann, *Diffusion flames*, Proc. Combust. Inst. 1 (1928), pp. 2–11.

[16] N. Peters, *Laminar diffusion flamelet models in non-premixed turbulent combustion*, Prog. Energy Combust. Sci. 10 (1984), pp. 319–339.

[17] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning*, Springer, New York , 2013.

[18] D. L. Banks, R. T. Olszewski, R. A. Maxion, *Comparing methods for multivariate nonparametric regression*, Comm. Stat. Sim. Comput. 32 (2003), pp. 541–571.

[19] J. H. Friedman, *Estimating functions of mixed ordinal and categorical variables using adaptive splines*, Technical Report, Dept. Statistics Stanford University (1991).

[20] J. H. Friedman, *Fitting functions to noisy data in high dimensions*, SLAC-PUB-4676 (1988).

[21] J. H. Friedman, B. W. Silverman, *Flexible parsimonious smoothing and additive modeling*, Technometrics 31 (1989), pp. 3–21.

[22] N. H. Qamar, Z. T. Alwahabi, Q. N. Chan, G. J. Nathan, D. Roekaerts, K. D. King, *Soot volume fraction in a piloted turbulent jet non-premixed flame of natural gas*, Combust. Flame 156 (2009), pp. 1339–1347.

[23] P. Billingsley, *Probability and Measure*, Wiley, New York, 1995.