

On Interpretability and Scheduling Space Choice for System Identification with Dynamic Takagi–Sugeno Models

Konrad Lange¹, Ralf Mikut², Andreas Kroll¹

¹Measurement and Control Research Group, University of Kassel
Mönchebergstr. 7, 34125 Kassel, Germany
E-Mail: {konrad.lange, andreas.kroll}@mrt.uni-kassel.de

²Institute for Automation and Applied Informatics (IAI), Karlsruhe Institute of
Technology (KIT)
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
E-Mail: ralf.mikut@kit.edu

Abstract

Dynamic Takagi–Sugeno models are often motivated by their interpretability. However, the choice of the scheduling variables with respect to interpretability is comparatively little discussed for auto-regressive models. In contrast to many neural network and kernel method approaches, Takagi–Sugeno models can expose where in the operating space the behavior changes. The scheduling variables directly influence how the nonlinear behavior can be approximated by affine local models. Here, interpretability primarily means that the partitioning of the scheduling space follows physical expectation, as opposed to other measures such as interpretability of the local models.

This paper examines how the choice of scheduling variables affects the interpretability and accuracy of auto-regressive Takagi–Sugeno models with cluster-based membership functions. Both a synthetic and a real-world example show that interpretability and accuracy are not necessarily a trade-off under a limited budget of scheduling dimensions and local models.

1 Introduction

Takagi–Sugeno (TS) models [1] achieve the approximation of nonlinear functions and systems by weighting the output of simpler local models by fuzzy basis functions. Unlike piecewise-affine (PWA) models, TS models with continuously differentiable basis functions provide smooth transitions between the local models [2]. Compared with many neural network approaches, TS models have an internal structure that can allow for controller design and better interpretation. Furthermore, established and interpretable methods can be applied to construct the linear or affine local models. Moreover, because their number of parameters is moderate, TS models can often be identified from less training data.

In the recent machine learning literature, a growing focus lies on interpretability [3–6]. The choice of TS models is often explained by interpretability of the models. Fuzzy systems, such as Mamdani systems, originally used rule bases with inherent linguistic interpretability incorporating expert knowledge [7]. TS models introduced data-driven identification of the rule consequents [1]. Although interpretability of fuzzy systems has been of interest for a considerable time [8, 9], it remains an active research topic [10].

Whereas some works focus on *local* interpretability [11–13], there is a multitude of interpretability measures [8, 9, 14]. Among others, a smaller number of inputs, linguistic terms and rules can be considered better interpretable [9]. Interpretability issues of TS models identified from data have been addressed [9, 12]. Many papers address interpretability and accuracy in classification [9, 15] or (static) regression with TS models [16]. Specifically, a trade-off is for example mentioned in [15, 16] and examples that implement multi-objective optimization approaches on the Pareto front of accuracy and parsimony or interpretability include [17–20].

In data-driven identification of dynamic discrete-time TS models, the scheduling variables are often chosen to be identical to, or as a subset of, the regressor variables (examples include [21]). This choice is not necessarily effective or physically interpretable when the nonlinear behavior depends on functions of the regressors. An example is nonlinear friction, which may depend on the

velocity rather than simply on the position alone. In this work, interpretability therefore refers to whether the partitioning follows physical expectation.

This work aims to show that choosing the scheduling variables based on physical insight can be beneficial. If the system's behavior depends on an operating condition such as high and low velocity, the identification procedure should (explicitly) incorporate this physically meaningful quantity for partitioning. Interpretable membership functions distinguish TS models within the broader class of mixture-of-experts (MoE) approaches [22]. Those approaches can also use linear submodels, but typically schedule them with a less interpretable gating network.

The idea of using physically meaningful scheduling variables is by no means new. Examples from the literature call this *semi-mechanistic fuzzy modeling* [23, 24]. In [25], a TS state-space model of a servo-pneumatic actuator is derived from a nonlinear physical model, with premise variables chosen based on the actuator physics. This contribution investigates in more depth how the choice of the scheduling space affects both the accuracy and the interpretability of the resulting TS model.

Related work uses principal component analysis (PCA) in the scheduling space of linear parameter-varying (LPV) models [26] and dynamic TS models [27]. Hierarchical local model trees (HILOMOT) are used for nonlinear regression in [28], where the split direction is based on the curvature of the residuals. This makes nonlinear effects interpretable along the split directions.

The rest of the paper is structured as follows: Section 2 introduces the specific TS model class considered here. Section 3 contains the main investigation and considerations of this work regarding the scheduling space choice. Test systems are introduced in Section 4, whereas the results are presented and discussed in Section 4.

2 Dynamic Takagi–Sugeno Models

This work is concerned with discrete-time local affine Takagi–Sugeno models originally introduced in [1]. Here, multi-dimensional membership functions are

used [2]. In general, the prediction $\hat{y}$ of TS models is a sum of the c local model predictions $\hat{y}_i$ weighted by the fuzzy basis functions ϕ_i . For the single-input single-output (SISO) models considered here, the prediction is:

$$\hat{y}(k) = \sum_{i=1}^c \phi_i(\mathbf{z}(k), \boldsymbol{\theta}_{\text{MF}}) \cdot \hat{y}_i(\mathbf{x}(k), \boldsymbol{\theta}_{\text{LM},i}). \quad (1)$$

Here, the local models are auto-regressive with exogenous inputs (ARX) or output error (OE) models [29], resulting in a TS model of nonlinear ARX (NARX) or nonlinear OE (NOE) type. The TS-NARX model is optimized with respect to the one-step-ahead (OSA) prediction error, whereas the TS-NOE model is optimized with respect to the (recursive) simulation error.

In the case of TS-NOE models, the local model predictions are:

$$\begin{aligned} \hat{y}_i(\mathbf{x}(k), \boldsymbol{\theta}_{\text{LM},i}) &= \mathbf{x}(k)^T \cdot \boldsymbol{\theta}_{\text{LM},i} \\ &= \begin{bmatrix} \hat{y}(k-1) & \cdots & \hat{y}(k-n) & u(k-\tau-1) & \cdots & u(k-\tau-m) & 1 \end{bmatrix} \\ &\quad \cdot \begin{bmatrix} -a_{1,i} & \cdots & -a_{n,i} & b_{1,i} & \cdots & b_{m,i} & c_i \end{bmatrix}^T. \end{aligned} \quad (2)$$

The regressor vector $\mathbf{x}(k) \in \mathbb{R}^p$ includes $m \in \mathbb{N}$ time-delayed inputs and $n \in \mathbb{N}$ time-delayed predicted outputs, with the discrete dead time τ . In the case of local ARX models, delayed measurements $y(k-j)$ are used instead of the predicted outputs $\hat{y}(k-j)$. The local parameter vectors are combined in $\boldsymbol{\theta}_{\text{LM}}^T = [\boldsymbol{\theta}_{\text{LM},1}^T \cdots \boldsymbol{\theta}_{\text{LM},c}^T]$.

The fuzzy basis functions are cluster-based membership functions of fuzzy c-means (FCM) [30] type, with the fuzziness parameter $\nu > 1, \nu \in \mathbb{R}$:

$$\phi_i(\mathbf{z}(k), \boldsymbol{\theta}_{\text{MF}}) = \mu_i(\mathbf{z}(k)) = \left[\sum_{j=1}^c \left(\frac{\|\mathbf{z}(k) - \mathbf{v}_i\|_2}{\|\mathbf{z}(k) - \mathbf{v}_j\|_2} \right)^{\frac{2}{\nu-1}} \right]^{-1}. \quad (3)$$

The variables $\mathbf{z}(k) \in \mathbb{R}^q$ used in determining the membership degree are called scheduling variables and span the scheduling space. The membership function parameters $\boldsymbol{\theta}_{\text{MF}}^T = [\boldsymbol{\theta}_{\text{MF},1}^T \cdots \boldsymbol{\theta}_{\text{MF},c}^T]$ are the locations of the c cluster centers $\mathbf{v}_j$

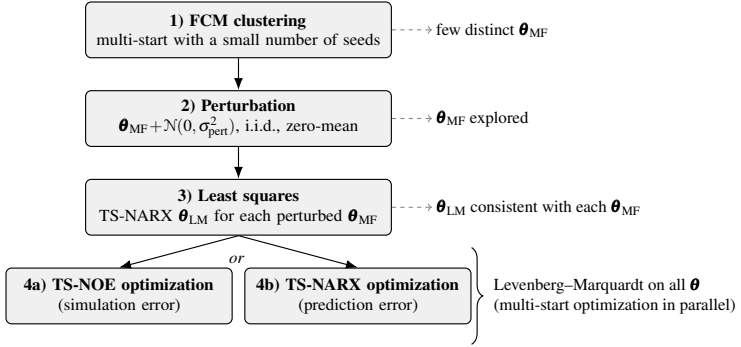


Figure 1: Overview of the multi-start optimization strategy to create initial parameter vectors through FCM clustering 1), perturbation of the cluster centers 2) and least-squares estimation of the local model parameters 3) for the subsequent gradient-based optimization 4a) or 4b).

in the scheduling space. The TS model parameter vector is therefore:

$$\boldsymbol{\theta}^T = \begin{bmatrix} \boldsymbol{\theta}_{MF}^T & \boldsymbol{\theta}_{LM}^T \end{bmatrix}. \quad (4)$$

The complete parameter vector $\boldsymbol{\theta}$ of the TS-NARX and TS-NOE models is optimized with the gradient-based Levenberg–Marquardt algorithm [31]. To reduce the risk of converging to poor local minima, a multi-start (MS) strategy with initial values obtained from TS-NARX estimates is used [32] (Fig. 1). First, $\boldsymbol{\theta}_{MF}$ is determined by FCM clustering with the Euclidean norm, using 8 random initializations of the clustering algorithm. To generate 128 MS initializations, the cluster centers are then perturbed by zero-mean, independent and identically distributed (i.i.d.) Gaussian noise with standard deviation σ_{pert} . For each perturbed $\boldsymbol{\theta}_{MF}$, the local model parameters $\boldsymbol{\theta}_{LM}$ are estimated globally by least squares. Starting from these initializations, all parameters $\boldsymbol{\theta}$ are optimized with respect to the prediction or simulation error. A separate validation dataset is used for early stopping and for selecting the best MS solution. The influence of the number of local models c for different choices of $\mathbf{z}$ is investigated in this work.

3 Scheduling Space Choice

3.1 Effect on Approximation of Nonlinearities

Similar to neural networks, TS models are often introduced as universal approximators. In practice, the theoretical conditions for universal approximation are usually not met, for example because the scheduling variables are only a subset of the regressors and the number of local models is limited. In general, universal approximation results [33, 34] only apply if all regressor variables $\mathbf{x}$ are part of the scheduling variables $\mathbf{z}$ or if the elements in $\mathbf{x}$ can be reconstructed from $\mathbf{z}$ (meaning there exists an injective mapping). This of course includes the case $\mathbf{z} = \mathbf{x}$.

For the sake of the following argument, the regressor vector defined in (2) is now split as

$$\mathbf{x}(k)^T = \begin{bmatrix} \mathbf{x}_N(k)^T & \mathbf{x}_L(k)^T \end{bmatrix}, \quad (5)$$

where $\mathbf{x}_N(k)$ is the subset of the regressors on which the system's output $y(k)$ depends nonlinearly, whereas the output only shows linear dependence on $\mathbf{x}_L(k)$. With $c \rightarrow \infty$, an arbitrarily fine clustering in the space spanned by the selected scheduling variables

$$\mathbf{z}(k) = \mathbf{x}_N(k) \quad (6)$$

enables universal approximation. In system identification practice, the regressors are often highly correlated and their full inclusion in the scheduling vector is not necessary [35].

Baseline results for well-known nonlinear benchmarks show that auto-regressive models often perform well with a relatively large number of lagged inputs and outputs [36]. In the study, the lag structures were selected based on the Akaike information criterion (AIC) evaluated with a linear ARX model and include a total of 17 to 23 regressors [36]. Using this number of scheduling dimensions would be neither effective nor interpretable. In applications, the number of local models is limited by computational resources. Moreover, the amount of training data required to avoid overfitting generally grows with model size [37].

This work shows that, instead of (6), it can be advantageous to select

$$\mathbf{z}(k) = \mathbf{W}\mathbf{x}_N(k) \quad (7)$$

to span the scheduling space based on information about the (physical) system. This can reduce the order if no strong dependence on the individual components of $\mathbf{x}_N(k)$ is observed. This can for example be the case when there are many regressors and/or the regressors are highly collinear. This approach could therefore reduce the curse of dimensionality that arises when a limited number of clusters is used in a high-dimensional clustering space. This benefit is conditional on the assumption that $\mathbf{W} \in \mathbb{R}^{q \times p}$ captures the directions along which the system is genuinely nonlinear. If this assumption is violated, reducing q introduces an irreducible error that does not vanish as $c \rightarrow \infty$. Even without dimension reduction, $\mathbf{W}$ can be beneficial if it aligns the cluster boundaries with the directions in which the dynamics change, so that fewer affine local models are needed. However, interpretability is only achieved when $\mathbf{W}$ has a small number of non-zero entries and is constructed using prior physical knowledge. Using a covariance matrix (or another data-driven estimate) to construct $\mathbf{W}$ does not automatically result in an interpretable scheduling space. Another way to improve interpretability is to merge compatible clusters [24]. However, this results in an accuracy–interpretability trade-off if many local models are needed to capture the direction of the nonlinear dependence.

3.2 Dependence on the Clustering Algorithm

Widely used clustering algorithms for dynamic TS models are fuzzy c-means (FCM) [30], Gustafson–Kessel (GK) [38], and Gath–Geva (GG) [39]. They differ mainly in their distance measure. FCM uses a fixed norm shared by all clusters. Usually this is either the Euclidean norm, which produces hyperspherical clusters, or the Mahalanobis norm, which produces hyperellipsoidal clusters of identical shape and orientation aligned with the principal axes of the data. GK instead adapts a cluster-specific norm based on the fuzzy covariance matrix, which allows arbitrarily oriented ellipsoidal clusters. GG additionally allows clusters of different volumes and densities.

The importance of an informed scheduling space depends on the clustering algorithm. It is particularly beneficial for FCM with the Euclidean distance norm, because this algorithm is sensitive to scaling. For example, because $y(k-1)$ and $y(k-2)$ are highly correlated, their difference has a small range, and FCM will not place clusters along this direction with the Euclidean distance norm. GK and GG are less affected because they can detect clusters elongated along hyperplanes [35]. However, both algorithms use more strategy parameters, are more likely to converge to local minima, and are more computationally expensive [40]. The ability to detect these clusters sacrifices axis-aligned interpretability and requires more parameters to be estimated. Furthermore, the clustering results are only used as initialization of the parameters θ_{MF} . Regardless of which clustering algorithm is used to obtain this initialization, the subsequent gradient-based optimization uses FCM-type membership functions here, and the dimensions of the scheduling space remain unchanged during optimization for the considered models. Therefore, without also adopting more flexible membership functions with corresponding parameters, any additional information gained from a more flexible clustering algorithm such as GK or GG would be lost in the subsequent TS model optimization. Accordingly, only FCM is used here, consistent with the chosen FCM-type membership functions.

4 Test Systems for Case Studies

4.1 Simulated Test System

The first example is a well-known test system proposed by Narendra and Parthasarathy [41]:

$$y(k) = \frac{y(k-1)y(k-2)(y(k-1)+2.5)}{1+y(k-1)^2+y(k-2)^2} + u(k-1). \quad (8)$$

The nonlinearity depends on $y(k-1)$ and $y(k-2)$, whereas $u(k-1)$ enters linearly. The system is excited with an amplitude-modulated pseudo-random binary signal (APRBS) and the resulting dataset is displayed in Fig. 2. Since the true system is known and regressor selection is not the focus of this study,

the regressors are set to $\mathbf{x}(k) = [y(k-1), y(k-2), u(k-1)]^T$. The fuzziness parameter $\nu = 1.3$ is chosen following [42]. The investigated choices for the scheduling space for the Narendra system are listed in Tab. 1.

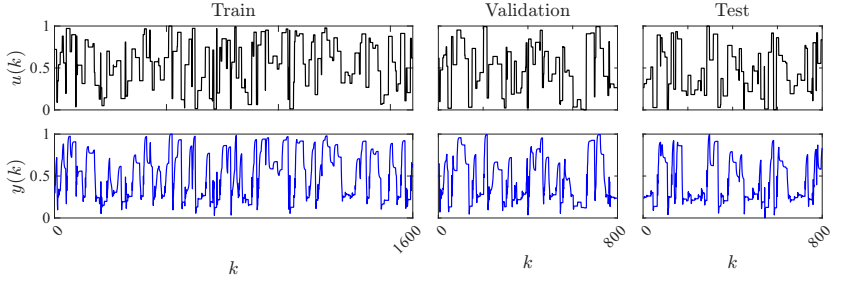


Figure 2: Scaled input and output signal of the Narendra system with the used APRBS excitation signal. The training, validation, and test datasets are shown.

4.2 Servo-Pneumatic Linear Drive

The servo-pneumatic linear drive (SPLD) and the dataset (Fig. 3) are described in detail in [32]. The input u is the voltage applied to the valve, while the position of the carriage is the measured output y . Both u and y are normalized to $[0, 1]$. Unlike the simulated system, the real-world data contains measurement noise and stochastic behavior caused by stick-slip friction. The system exhibits nonlinear behavior due to several sources, including:

- valve characteristics depending on u ,
- compressibility of air and chamber thermodynamics depending on y ,
- and the friction depending on y and $\dot{y}$.

Therefore, an informed choice of the scheduling space includes $u(k-\tau-1)$, $y(k-1)$ as well as $y(k-1) - y(k-2)$ as an approximation of the velocity $\dot{y}$ at time step $k-1$. The difference quotient $(y(k-1) - y(k-2))/T_s$ differs from the plain difference only by the constant $1/T_s$, which is absorbed by the normalization of the $\mathbf{z}(k)$ to $[0, 1]$.

The TS model structure was taken from an investigation on the same dataset, where it was selected from the Pareto front of the validation MSE and the number of model parameters [32]. The regressors are $\mathbf{x}(k) = [y(k-1), \dots, y(k-4), u(k-\tau-1), \dots, u(k-\tau-6), 1]^T$, with $\tau = 2$, and the fuzziness parameter is chosen as $\nu = 1.3$.

The different choices for the scheduling space for the SPLD system are listed in Tab. 1. The reduced variants are chosen as minimal examples: one including only $[y(k-1), u(k-3)]$ and one additionally including $y(k-2)$. This allows evaluation of whether the additional information of $y(k-2)$ is beneficial. The input $u(k-3)$ is included in all scheduling space choices, as a dependence of the nonlinear behavior on the input cannot be ruled out for the uninformed variants, and is known to exist for the physically informed choice. Further variants could be considered, but the selection is kept concise to maintain the clarity of the study.

5 Results

5.1 Accuracy

Narendra System:

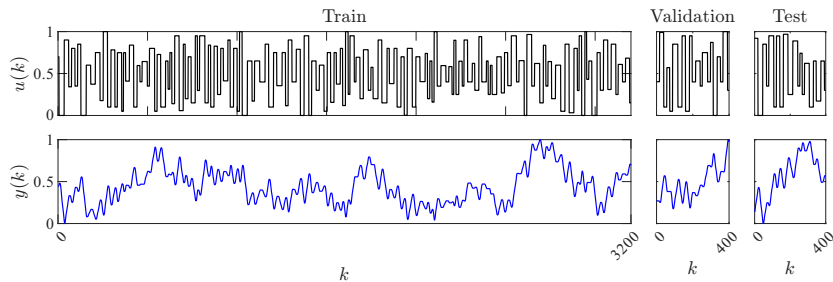


Figure 3: Scaled input and output signal of the servo-pneumatic linear drive (SPLD). The training, validation, and test datasets are shown.

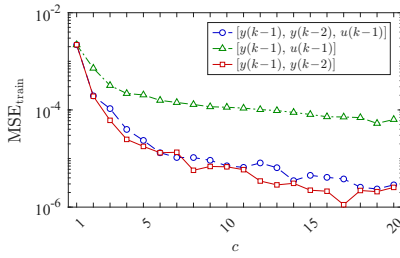
Variant	Scheduling space $\mathbf{z}$
<i>Narendra system</i>	
full regressors	$[y(k-1), y(k-2), u(k-1)]$
reduced regressors	$[y(k-1), u(k-1)]$
informed	$[y(k-1), y(k-2)]$
<i>Servo-pneumatic linear drive (SPLD)</i>	
full regressors	$[y(k-1), \dots, y(k-4), u(k-3), \dots, u(k-8)]$
reduced regressors	$[y(k-1), y(k-2), u(k-3)]$
reduced regressors	$[y(k-1), u(k-3)]$
informed, without $y(k-1)$	$[y(k-1) - y(k-2), u(k-3)]$
informed	$[y(k-1), y(k-1) - y(k-2), u(k-3)]$

Table 1: Investigated scheduling space choices.

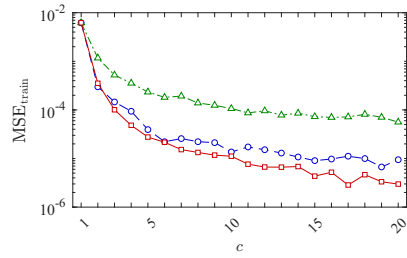
For the Narendra system, a linear model ($c = 1$) cannot adequately approximate the nonlinear behavior, and all TS models with $c > 1$ perform significantly better (Fig. 4). The OSA prediction and simulation errors for increasing c support choosing the scheduling space according to the known nonlinear dependencies of the system. Omitting $y(k-2)$ degrades the accuracy significantly, whereas omitting $u(k-1)$ improves it. This holds for all investigated numbers of local models. As expected for noise-free data, the TS-NOE simulation results are similar to the TS-NARX prediction results.

SPLD:

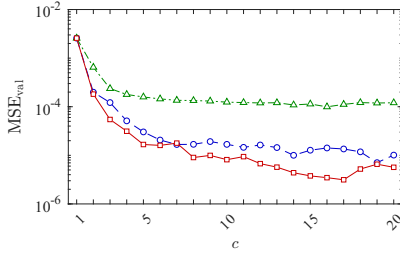
The results of the different scheduling space choices for the SPLD are shown in Fig. 5. For OSA prediction, the MSE is very small for all scheduling spaces, but the informed choice still performs best. Including only the difference $y(k-1) - y(k-2)$ and omitting $y(k-1)$ performs similarly, except on the validation dataset. Unlike the Narendra system, the TS-NOE simulation results differ considerably from the TS-NARX results. TS models with a moderate number of local models and a suitable $\mathbf{z}$ outperform the linear OE model, which again indicates nonlinear behavior. This is expected, as the TS-NARX models are trained and evaluated on the prediction error using measured past outputs, so their errors do not accumulate. The TS-NOE test error is therefore the



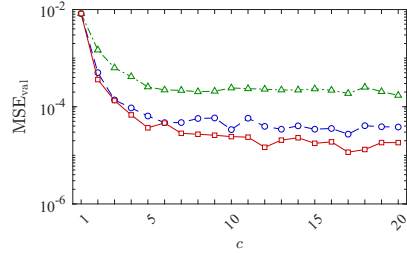
(a) TS-NARX, prediction MSE, training dataset.



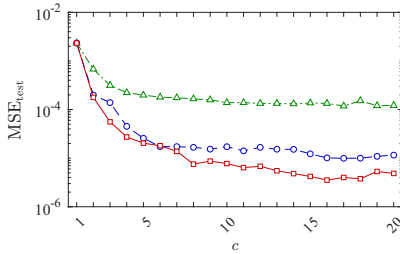
(b) TS-NOE, simulation MSE, training dataset.



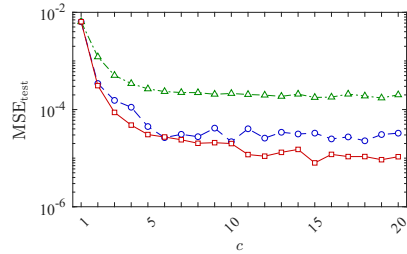
(c) TS-NARX, prediction MSE, validation dataset.



(d) TS-NOE, simulation MSE, validation dataset.



(e) TS-NARX, prediction MSE, test dataset.



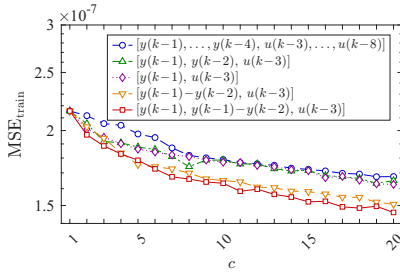
(f) TS-NOE, simulation MSE, test dataset.

Figure 4: Prediction and simulation error (Narendra system) depending on the choice of the scheduling space and the number of local models c . Left column: TS-NARX (prediction-error optimized), one-step-ahead prediction MSE. Right column: TS-NOE (simulation-error optimized), simulation MSE. Rows: training (top), validation (middle) and test (bottom) datasets. The informed choice of the scheduling space is shown in red.

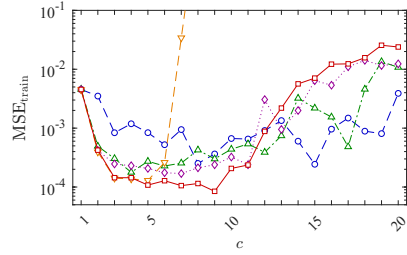
more realistic indicator of model quality for practical deployment, since it reflects performance in a free-run simulation. Using all regressors as scheduling variables is generally not advisable. Without explicitly including $y(k-1) - y(k-2)$, the information of $y(k-2)$ is not beneficial to the model accuracy. For small c , the physically informed choice is clearly beneficial. In contrast to the simulated case study, the results of the MSE on the training and test datasets differ significantly and are less clear for increasing numbers of local models.

On the training data, including $y(k-1) - y(k-2)$ clearly improves performance for moderate c . In contrast, adding $y(k-2)$ itself to $[y(k-1), u(k-3)]$ has no significant effect. This supports the hypothesis that the informed choice improves accuracy by separating the nonlinear behavior along this direction. Up to $c = 8$ local models, the test results are mostly consistent with the training results. For higher numbers of local models, the results can only be interpreted as a rough trend. An explanation could be that the amount of training data per local model decreases as more local models are used. This also applies to the data available for the transitions between local models, which can lead to overfitting. The problem is further aggravated by the stochastic system behavior caused by the stick-slip effect.

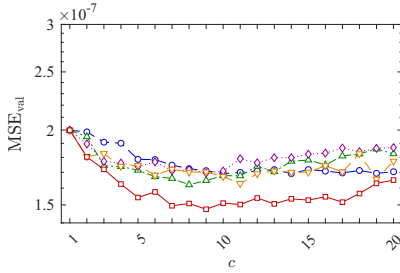
With more local models, the cluster centers lie closer together and each cluster spans a smaller region of high membership degree. Since only predicted outputs are used in $\mathbf{z}(k)$, the scheduling variables deviate from their values based on measured outputs. An error of the same magnitude is therefore more likely to shift $\mathbf{z}(k)$ into the region of a neighboring cluster, activating a local model identified for a different operating region. This applies in particular to the difference $\hat{y}(k-1) - \hat{y}(k-2)$, which makes the scheduling more susceptible to errors. The scheduling variables $\mathbf{z}$ are normalized, and the lagged outputs $y(k-j)$ are highly correlated. Consequently, even small errors in the model outputs $\hat{y}(k)$ can lead to the activation of different local models. In simulation, this effect is amplified by error propagation. This explanation is supported by the results for scheduling with $[y(k-1) - y(k-2), u(k-3)]$, where the absolute output information is lost. For this choice, competitive results for small c are followed by a drastic increase in MSE.



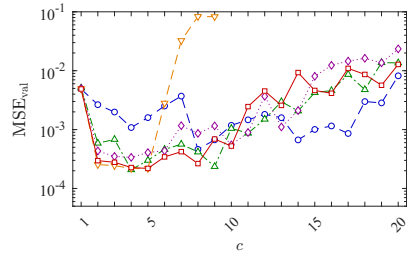
(a) TS-NARX, prediction MSE, training dataset.



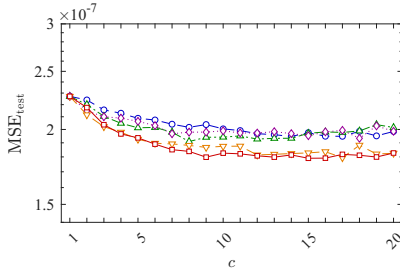
(b) TS-NOE, simulation MSE, training dataset.



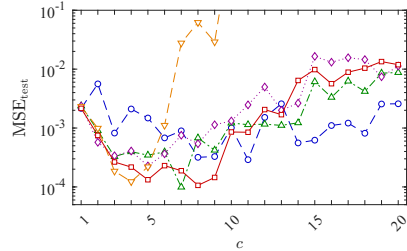
(c) TS-NARX, prediction MSE, validation dataset.



(d) TS-NOE, simulation MSE, validation dataset.



(e) TS-NARX, prediction MSE, test dataset.



(f) TS-NOE, simulation MSE, test dataset.

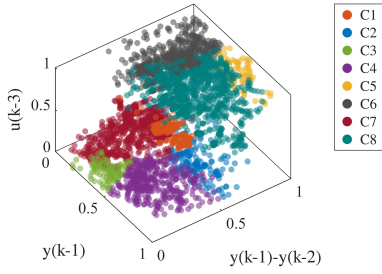
Figure 5: Prediction and simulation error (SPLD) depending on the choice of the scheduling space and the number of local models c . Left column: TS-NARX (prediction-error optimized), one-step-ahead prediction MSE. Right column: TS-NOE (simulation-error optimized), simulation MSE. Rows: training (top), validation (middle) and test (bottom) datasets. The physically informed choice of the scheduling space is shown in red.

Complete results for Akaike’s Information Criterion (AIC) and the Bayesian Information Criterion (BIC), which weigh model accuracy against model complexity [35], are not reported here, as a meaningful comparison would also require varying the number of regressors used by each local model. Preliminary results on the SPLD training data, obtained by varying the scheduling variables and the number of local models while keeping the regressor variables fixed, indicate that the physically informed choice yields lower AIC and BIC values.

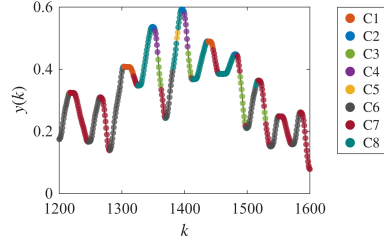
5.2 Interpretability

To assess interpretability, the membership degrees are visualized both in the scheduling space and over a section of the time series. Each sample is colored according to its highest membership degree. This loses interpretation of the transitions but improves visibility. For the uninformed choice (Fig. 6c), the data lies close to a diagonal because $y(k-1)$ and $y(k-2)$ are strongly correlated. The velocity enters the scheduling only implicitly, if at all. Using $y(k-1) - y(k-2)$ as a velocity proxy spreads the data more widely in the scheduling space, and the resulting partitioning persists after gradient-based optimization.

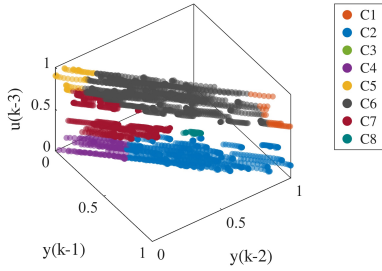
Because the difference makes changes in position explicit, the clusters partition the data along this direction. Fig. 6a depicts differences in clusters along the direction of the velocity (e.g., $C3$ and $C5$) that are not only due to correlation with the input. The clusters in Fig. 6a can be assigned to physical operating regions. For example, $C1$ corresponds to a slowly moving carriage near the center of the stroke with a small absolute input (as positive/negative u are normalized to $[0, 1]$). In contrast, $C5$ represents fast motion with a large input, largely independent of the absolute position. The resulting difference is also visible when the fuzzy membership to the clusters is highlighted in the time series. Fig. 6b shows much clearer separation based on velocity than Fig. 6d and Fig. 6f. The apparent velocity-based differentiation in Fig. 6d and Fig. 6f is largely explained by the correlation between $u(k-3)$ and $y(k-1) - y(k-2)$ at most data points, rather than by explicitly separating different velocities.



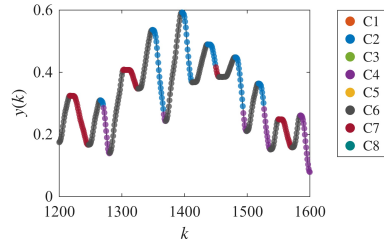
(a) Informed choice.



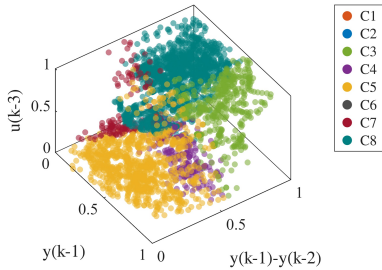
(b) Informed choice.



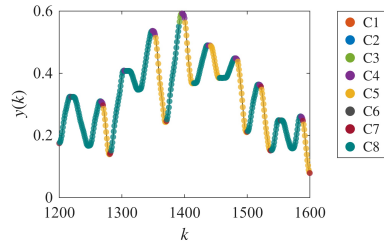
(c) Uninformed choice (reduced regressors).



(d) Uninformed choice (reduced regressors).



(e) Uninformed choice (all regressors used): membership projected to physically informed dimensions.



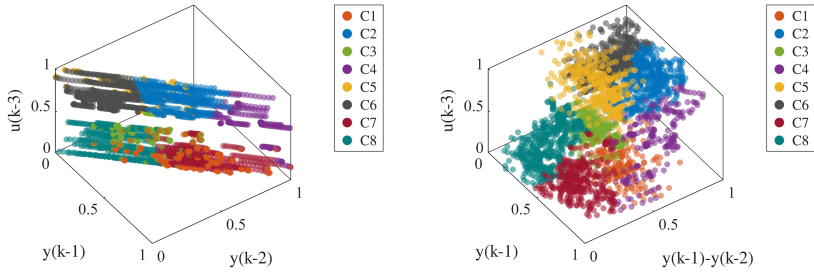
(f) Uninformed choice (all regressors used).

Figure 6: Comparison of the physically informed and uninformed choice of scheduling variables for the SPLD case study. Initial clustering is carried out with the FCM algorithm and the Euclidean distance norm after normalization of the input features. The membership shown is the result after gradient-based optimization for TS-NOE. Eight local affine models are used and the highest fuzzy membership degree to the clusters C_i is depicted by crisp colors. The results are displayed in the scheduling space (left) and over a part of the time series (right).

5.3 Comparison to Clustering with the Mahalanobis Norm

FCM clustering with the Euclidean norm is sensitive to scaling and restricted to hyperspherical clusters. In contrast, FCM clustering with the Mahalanobis norm can detect clusters that are aligned with the principal directions of the data and have a hyperellipsoidal shape [40]. This is achieved by calculating the covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{q \times q}$ of the data in the scheduling space. To preserve the cluster shape information determined from the data scattering during TS model optimization, the distance norm in (3) has to be changed to the Mahalanobis norm. The inverse $\mathbf{\Sigma}^{-1}$ of the covariance matrix is used as the form matrix. In the simplest case $\mathbf{\Sigma}$ is kept fixed, and the number of model parameters to be optimized with respect to prediction or simulation error remains the same as with the Euclidean norm. It increases slightly if a shared matrix $\mathbf{\Sigma}$ is optimized, but considerably if the shape of each cluster is optimized individually during gradient-based TS optimization. Preliminary results for the SPLD dataset using a fixed matrix $\mathbf{\Sigma}$ indicate a numerical problem when optimizing for the simulation error. The large eigenvalue of $\mathbf{\Sigma}^{-1}$ associated with the difference direction amplifies errors in $\hat{y}(k-1) - \hat{y}(k-2)$ direction. These errors then propagate recursively through the simulation. Strongly regularizing $\mathbf{\Sigma}$ mitigates this amplification. However, it also removes the separation into clusters along $y(k-1) - y(k-2)$. Figure 7 therefore shows the result after FCM clustering with the Mahalanobis norm (without a subsequent gradient-based TS optimization with respect to the model prediction error).

In the coordinates of the scheduling space (Fig. 7a), several clusters overlap at a similar position and input, e.g., $C1$ and $C7$ or $C3$ and $C8$. Projected onto the physically informed dimensions (Fig. 7b), the same clusters form regions that are clearly separated along $y(k-1) - y(k-2)$. This is consistent with Section 3: since $y(k-1)$ and $y(k-2)$ are highly correlated, one principal axis of $\mathbf{\Sigma}$ approximately corresponds to their difference, and the Mahalanobis norm amplifies this low-variance direction. Through clustering with the Mahalanobis norm, the scaling that the physically informed choice achieves through the normalization of $y(k-1) - y(k-2)$ is determined from data. However, two limitations remain apart from the described problem in simulation error optimization. First, the partitioning becomes interpretable only after projection onto the physically



(a) FCM clustering results (Mahalanobis norm), displayed in scheduling space. (b) FCM clustering results (Mahalanobis norm) projected to physically informed dimensions.

Figure 7: Fuzzy membership degree using $[y(k-1), y(k-2), u(k-3)]$ as the scheduling space on the SPLD data. FCM clustering results with the Mahalanobis norm are shown without gradient-based TS-NOE optimization. The membership is shown in the scheduling space's coordinates (left) and projected onto the physically informed dimensions $[y(k-1), y(k-1)-y(k-2), u(k-3)]$ (right). The highest membership degree to each cluster C_i is depicted by crisp colors; 8 clusters are used.

informed dimensions. Second, the Mahalanobis norm weights the low-variance directions, regardless of whether the system is nonlinear along them.

5.4 Discussion

For the limited case studies with two systems, the physically informed and therefore more easily interpretable scheduling space can also result in better model accuracy. Interestingly, while the prediction and simulation errors decrease steadily for the simulated example without noise (Fig. 4), there is an increase for $c > 8$ local models for the SPLD example. Here, the more accurate models are more interpretable as well. For example, $c = 8$ fuzzy clusters in a 3-dimensional scheduling space can be interpreted visually (Fig. 6) because the dimensions allow for mapping to system behavior. Analyzing a 3-dimensional scheduling space also makes it possible to identify clusters with a small number of training data points. The steeper MSE increase with $c > 8$ for TS-NOE in Fig. 5 could be explained by the fact that errors of past predicted outputs propagate into the membership degrees.

The use of the Euclidean distance norm is closely related to the observed improvement. Possible improvements when using more geometrically flexible distance norms, such as the Mahalanobis norm, must be investigated with different membership functions to preserve the benefit during gradient-based TS-NOE optimization. Although the fixed fuzziness parameter $v = 1.3$ for all clusters does not appear to limit the model performance here, it reduces the generality of the observations. Other values of v could be examined in future work. It is important to note that this contribution does not argue that including physical knowledge is necessary to achieve accuracy in data-driven system identification with the considered modeling approach. Other contributions, such as the work [43], show that the automatic approximation of the nonlinearity by tree-based local model networks can deliver competitive performance for benchmark problems.

6 Conclusion and Outlook

Interpretability and accuracy of dynamic cluster-based TS models depend heavily on an adequate choice of the scheduling space, and they do not always form a trade-off. With a limited number of scheduling dimensions and local models, it is crucial to choose physically interpretable features on which the output depends nonlinearly, so that the model can effectively approximate the nonlinear dynamic behavior. Results from two case studies support this. If the true nonlinearity is not fully captured by the chosen scheduling space, an irreducible bias is introduced that no increase in the number of local models can remove. This is unlike fully data-driven alternatives, which (in principle, given enough data and successful optimization) could still discover the missing direction when all regressors are used.

In future work, the investigations have to be extended to a broader range of test systems and benchmarks. It is also necessary to investigate gradient-based parameter optimization for simulation error with membership functions with flexible cluster shapes.

7 Acknowledgment

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 541311230.

References

- [1] T. Takagi and M. Sugeno, “Fuzzy identification of systems and its applications to modeling and control,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-15, no. 1, pp. 116–132, 1985.
- [2] A. Kroll, “Identification of functional fuzzy models using multidimensional reference fuzzy sets,” *Fuzzy Sets and Systems*, vol. 80, no. 2, pp. 149–158, 1996.
- [3] D. Materassi, S. Warnick, C. Rojas, M. Schoukens, and E. Cross, “Explaining complex systems: a tutorial on transparency and interpretability in machine learning models (Part I),” in *20th IFAC Symposium on System Identification (SYSID 2024)*, vol. 58, no. 15, 2024, pp. 492–496.
- [4] Y. Zhang, P. Tiño, A. Leonardis, and K. Tang, “A Survey on Neural Network Interpretability,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 5, pp. 726–742, 2021.
- [5] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A Survey of Methods for Explaining Black Box Models,” *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.
- [6] K. Zhang, T. Shao, Y. Sun, X. Xu, X. Zhang, X. Zhou, K. Ding, and S. Huang, “Interpretable research of fuzzy methods: A literature survey,” *Information Fusion*, vol. 126, p. 103524, 2026.
- [7] E. H. Mamdani and S. Assilian, “An experiment in linguistic synthesis with a fuzzy logic controller,” *International Journal of Man-Machine Studies*, vol. 7, no. 1, pp. 1–13, 1975.

- [8] S. Guillaume, “Designing fuzzy inference systems from data: An interpretability-oriented review,” *IEEE Transactions on Fuzzy Systems*, vol. 9, no. 3, pp. 426–443, 2001.
- [9] R. Mikut, J. Jäkel, and L. Gröll, “Interpretability issues in data-based learning of fuzzy systems,” *Fuzzy Sets and Systems*, vol. 150, no. 2, pp. 179–197, 2005.
- [10] L. Pickering, K. Cohen, and B. De Baets, “A Narrative Review on the Interpretability of Fuzzy Rule-Based Models from a Modern Interpretable Machine Learning Perspective,” *International Journal of Fuzzy Systems*, vol. 28, no. 4, pp. 1313–1332, 2026.
- [11] J. Yen, L. Wang, and C. W. Gillespie, “Improving the interpretability of TSK fuzzy models by combining global learning and local learning,” *IEEE Transactions on Fuzzy Systems*, vol. 6, no. 4, pp. 530–537, 1998.
- [12] T. A. Johansen, R. Shorten, and R. Murray-Smith, “On the interpretation and identification of dynamic Takagi-Sugeno fuzzy models,” *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 3, pp. 297–313, 2000.
- [13] J. Abonyi and R. Babuska, “Local and global identification and interpretation of parameters in Takagi-Sugeno fuzzy models,” in *Ninth IEEE International Conference on Fuzzy Systems. FUZZ- IEEE 2000*, vol. 2, 2000, pp. 835–840.
- [14] M. J. Gacto, R. Alcalá, and F. Herrera, “Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures,” *Information Sciences*, vol. 181, no. 20, pp. 4340–4360, 2011.
- [15] D. Nauck and R. Kruse, “How the learning of rule weights affects the interpretability of fuzzy systems,” in *1998 IEEE International Conference on Fuzzy Systems Proceedings. IEEE World Congress on Computational Intelligence*, vol. 2, 1998, pp. 1235–1240.
- [16] M. Setnes, R. Babuska, and H. B. Verbruggen, “Rule-based modeling: precision and transparency,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 28, no. 1, pp. 165–169, 1998.

- [17] T. A. Johansen and R. Babuska, “Multiobjective identification of Takagi-Sugeno fuzzy models,” *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 6, pp. 847–860, 2003.
- [18] M. J. Gacto, R. Alcalá, and F. Herrera, “Integration of an Index to Preserve the Semantic Interpretability in the Multiobjective Evolutionary Rule Selection and Tuning of Linguistic Fuzzy Systems,” *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 3, pp. 515–531, 2010.
- [19] C. Fuchs, U. Kaymak, and M. S. Nobile, “Building Interpretable and Parsimonious Fuzzy Models using a Multi-Objective Approach,” in *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2022, pp. 1–8.
- [20] Q. Chang, Y. Chen, J. Liang, C. Liu, and J. Wang, “Constructing Concise Instance-Based Takagi–Sugeno–Kang Fuzzy Systems via Multiobjective Particle Swarm Optimization,” *IEEE Transactions on Fuzzy Systems*, vol. 33, no. 11, pp. 4111–4123, 2025.
- [21] D. Aleksovski, D. Dovžan, S. Džeroski, and J. Kocijan, “A comparison of fuzzy identification methods on benchmark datasets,” in *4th IFAC Conference on Intelligent Control and Automation Sciences (ICONS 2016)*, vol. 49, no. 5, 2016, pp. 31–36.
- [22] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, “Adaptive Mixtures of Local Experts,” *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.
- [23] R. Babuska, *Fuzzy Modeling for Control*. Dordrecht: Springer, 1998.
- [24] J. Abonyi and B. Feil, “Fuzzy Clustering for System Identification,” in *Cluster Analysis for Data Mining and System Identification*. Birkhäuser Basel, 2007, pp. 141–224.
- [25] H. Schulte and H. Hahn, “Fuzzy state feedback gain scheduling control of servo-pneumatic actuators,” *Control Engineering Practice*, vol. 12, no. 5, pp. 639–650, 2004.

- [26] A. Kwiatkowski and H. Werner, “PCA-Based Parameter Set Mappings for LPV Models With Fewer Parameters and Less Overbounding,” *IEEE Transactions on Control Systems Technology*, vol. 16, no. 4, pp. 781–788, 2008.
- [27] D. Dovžan and I. Škrjanc, “Fuzzy Space Partitioning Based on Hyperplanes Defined by Eigenvectors for Takagi-Sugeno Fuzzy Model Identification,” *IEEE Transactions on Industrial Electronics*, vol. 67, no. 6, pp. 5144–5153, 2020.
- [28] M. Schöne, M. Kohlhasse, and O. Nelles, “Incorporation of structural properties of the response surface into oblique model trees,” *at - Automatisierungstechnik*, vol. 73, no. 10, pp. 727–739, 2025.
- [29] L. Ljung, *System identification: theory for the user*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.
- [30] J. C. Bezdek, R. Ehrlich, and W. Full, “FCM: The fuzzy c-means clustering algorithm,” *Computers & Geosciences*, vol. 10, no. 2, pp. 191–203, 1984.
- [31] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York, NY: Springer, 2006.
- [32] K. Lange and A. Kroll, “Multi-Start Initialization for Simulation Error Optimization of Takagi–Sugeno Model Parameters,” in *30th International Conference on System Theory, Control and Computing (ICSTCC)*, 2026, accepted.
- [33] B. Kosko, “Fuzzy systems as universal approximators,” *IEEE Transactions on Computers*, vol. 43, no. 11, pp. 1329–1333, 1994.
- [34] H. Ying, “Sufficient conditions on uniform approximation of multivariate functions by general Takagi-Sugeno fuzzy systems with linear rule consequent,” *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 28, no. 4, pp. 515–520, 1998.

- [35] O. Nelles, *Nonlinear System Identification: From Classical Approaches to Neural Networks, Fuzzy Models, and Gaussian Processes*, 2nd ed. Cham: Springer, 2020.
- [36] M. Champneys, G. I. Beintema, R. Tóth, M. Schoukens, and T. J. Rogers, “Baseline Results for Selected Nonlinear System Identification Benchmarks,” in *20th IFAC Symposium on System Identification (SYSID 2024)*, vol. 58, no. 15, 2024, pp. 474–479.
- [37] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer, 2006.
- [38] D. E. Gustafson and W. C. Kessel, “Fuzzy clustering with a fuzzy covariance matrix,” in *1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, 1978, pp. 761–766.
- [39] I. Gath and A. B. Geva, “Unsupervised optimal fuzzy clustering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 773–780, 1989.
- [40] A. Kroll, *Computational Intelligence: Probleme, Methoden und technische Anwendungen*, 2nd ed. Berlin: De Gruyter Oldenbourg, 2016.
- [41] K. S. Narendra and K. Parthasarathy, “Identification and control of dynamical systems using neural networks,” *IEEE Transactions on Neural Networks*, vol. 1, no. 1, pp. 4–27, 1990.
- [42] A. Kroll, “On choosing the fuzziness parameter for identifying TS models with multidimensional membership functions,” *Journal of Artificial Intelligence and Soft Computing Research*, vol. 1, no. 4, pp. 283–300, 2011.
- [43] J. Belz, T. Munker, T. O. Heinz, G. Kampmann, and O. Nelles, “Automatic Modeling with Local Model Networks for Benchmark Processes,” in *20th IFAC World Congress*, vol. 50, no. 1, 2017, pp. 470–475.