Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Uncertainty-aware reduced-order modelling with sparse sensing and data assimilation for forecasting combustion dynamics

Constanza Lopez
Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire EM2C
Alberto Procacci
Aero-Thermo-Mechanics Laboratory (ATM), Université Libre de Bruxelles, École Polytechnique de Bruxelles
Bérengère Podvin
Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire EM2C
Salvatore Iavarone
salvatore.iavarone@centralesupelec.fr

Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire EM2C

Method Article

Keywords:

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11381183/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Uncertainty-aware reduced-order modelling with sparse sensing and data assimilation for forecasting combustion dynamics

Constanza López[1], Alberto Procacci[2,3], Bérengère Podvin[1], Salvatore Iavarone[1*]

[1*]Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire EM2C, 91190, Gif-sur-Yvette, France.
[2]Aero-Thermo-Mechanics Department, Université Libre de Bruxelles, Avenue Franklin Roosevelt 50, Brussels, 1050, Belgium.
[3]Brussels Institute for Thermal-Fluid Systems and Clean Energy (BRITE), Université Libre de Bruxelles and Vrije Universiteit Brussel, Brussels, 1050, Belgium.

*Corresponding author(s). E-mail(s): salvatore.iavarone@centralesupelec.fr;
Contributing authors: constanza.lopez-lozano@centralesupelec.fr; alberto.procacci@ulb.be; berengere.podvin@centralesupelec.fr;

Abstract

Accurate and efficient prediction of combustion dynamics remains challenging due to the nonlinear and multiscale nature of reactive flows. High-fidelity simulations are often impractical for real-time applications due to their computational cost. Reduced-order models offer a promising alternative but their predictive capability may decay when the operating conditions differ from those represented in the training data. This work develops an uncertainty-aware reduced-order modelling framework for combustion dynamics, combining Multi-scale Proper Orthogonal Decomposition (mPOD), Gaussian Process Regression (GPR), sparse sensing and data assimilation. The latent dynamics extracted by mPOD is learned offline using GPR, while sparse sensing is employed during the forecast to enable data assimilation in the latent space from limited observations. The frequency organization introduced by mPOD and the characteristics temporal scales identified in the latent dynamics are used to define the temporal covariance structure of the GPR. Different GPR kernel formulations, spatial sensing

constraints, latent-state reconstruction methods, and assimilation frequencies are systematically assessed. The approach is evaluated on two different datasets: an experimental dataset comprising chemiluminescence images from a forced swirled turbulent flame, and a numerical dataset of an axi-symmetric, time-varying, non-premixed laminar flame. The results show that a Periodic GPR kernel provides the most robust latent forecast across both datasets, while regularized least squares and Bayesian sparse reconstruction methods improve robustness under limited spatial accessibility. Data assimilation further compensates for forecast deviations by combining noisy sparse observations with the probabilistic GPR forecast. In the laminar-flame case, six temperature measurements are sufficient to substantially enhance an inaccurate Cosine GPR kernel forecast and achieve prediction errors approaching those obtained with the best-performing Periodic kernel. These results demonstrates the potential of the proposal framework for uncertainty-aware forecasting of combustion dynamics by combining probabilistic reduced-order modelling, sparse observations and data assimilation.

Keywords: Combustion dynamics, Reduced-order modelling; Gaussian Process Regression, Sparse sensing, Data assimilation, Multi-scale Proper Orthogonal Decomposition

1 Introduction

Digital twins are emerging as a promising technology for monitoring and predicting the behavior of complex engineering systems by combining computational models with information from the physical system [1]. In practice, however, real-time measurements are often limited by accessibility, operating conditions, and instrumentation constraints. In combustion system, digital twins application is challenging due to the nonlinear and multi-scale nature of reacting flows and the computational cost of high-fidelity simulations [2, 3]. Reduced-order models (ROMs) have emerged as an attractive alternative for developing digital twins, as they enable the representation of the dominant system dynamics in a low-dimensional space while retaining the main features of the original system [4].

Modal decomposition techniques, such as Proper Orthogonal Decomposition (POD) and Dynamic Mode Decomposition (DMD) have been widely employed for this purpose [5, 6]. Recently, non-intrusive ROMs have combined modal decomposition with data-driven regression techniques to infer the evolution of the reduced-order state without explicitly projecting the governing equations [7, 8]. In combustion, POD-based representations have been coupled with surrogate models such as Kriging and Gaussian Process Regression (GPR) to reconstruct thermo-chemical fields at operating conditions not included in the training data. Applications include laminar flames configurations [9] and industrial-scale combustion systems [10]. However, classical POD may mix structures associated with different temporal scales, while DMD-based representations can be restrictive for strongly nonlinear or multi-frequency dynamics.

Multi-scale Proper Orthogonal Decomposition (mPOD) overcomes this by introducing frequency-band separation during the decomposition, providing latent trajectories with a more clearly defined temporal content [11].

In addition to parametric reconstruction, data-driven ROMs have been developed to predict combustion states and reduced dynamics beyond the conditions used for model construction. Swischuk *et al.* [12] developed predictive ROMs in transformed POD coordinates for a single-injector combustor, obtaining accurate predictions even outside the training interval. Ni *et al.* [13] investigated different regression strategies for mapping operating parameters to POD coefficients in multi-species mixing and combustion, highlighting the influence of the surrogate formulation on ROM predictive performance.

For time-dependent forecasting, the temporal evolution of the latent coefficients must be modelled explicitly. GPR is well suited to this task because its formulation provides probabilistic predictions and associated uncertainty [7, 14]. In the present work, the temporal covariance structure of the GPR is informed by the frequency separation introduced by mPOD and the characteristic frequencies identified in the latent dynamics. Different kernels are therefore compared to assess their ability to represent the oscillatory combustion dynamics.

In combustion systems, full flow or thermo-chemical states are rarely accessible experimentally. This means that reduced-order coefficients often need to be inferred from limited measurements. Sparse sensing approaches exploit the low-dimensional structure of the system to reconstruct the reduced state from a small set of spatial observations. The resulting accuracy depends on sensor placement, measurement noise and the conditioning of the inverse problem [15]. POD-based sparse reconstruction and optimized sensor placement are becoming broadly used to estimate flow states. In combustion, related approaches have been employed to recover flame and thermo-chemical information from limited measurements. For instance, Lopez *et al.*[16] reconstructed temperature and velocity fields from sparse temperature measurements, and Rosati *et al.* [17] presents a reduced-order reconstruction of a cyclonic combustor under constrained sensing configurations.

Although sparse reconstruction may estimate the reduced state, does not correct errors in the forecasting model. Sequential data assimilation aims to mitigate this issue by combining model predictions with new observations according to their uncertainties [18]. Kalman-filter-based approaches, particularly the Ensemble Kalman Filter (EnKF) introduced by Evensen [19], have been applied at different modelling levels in reacting flow problems. At the high-fidelity level, experimental velocity and flame measurements have been assimilated into large eddy simulations to improve the prediction of transient extinction and reignition dynamics [20]. Its potential for combustion digital twins has already been demonstrated by Donato *et al.*, applying EnKF-based data assimilation to update ROMs of a semi-industrial furnace using experimental measurements [21, 22].

While forecasting, sparse sensing, and data assimilation have each been investigated in reduced-order combustion modelling, fewer studies have considered their integration within a common uncertainty-aware framework. Procacci *et al.* [23] combined mPOD, GPR, sparse sensing, and EnKF-based data assimilation for stochastic

forecasting and correction of noisy combustion dynamics. This demonstrated that sparse observations can be assimilated directly in the reduced space to correct probabilistic forecasts. However, the interaction between forecasting model formulation, spatial sensing constraints, latent state reconstruction, measurement uncertainty, and assimilation frequency has not been systematically studied.

Motivated by these considerations, the present work extends the stochastic ROM framework introduced in [23] through a systematic assessment of frequency-informed GPR kernels, sparse-sensing configurations, latent-state reconstruction methods, and data-assimilation frequency. Particular attention is given to the propagation of uncertainty and the role of forecast and observation uncertainties in the reduced-order prediction and correction process. The methodology is evaluated on two complementary test cases: an experimental dataset of chemiluminescence images of a forced swirled flame, used to investigate spatial sensing constraints and assimilation frequency, and a numerical pulsated laminar flame, where temperature values sampled at optimally selected sensor locations are assimilated to improve a prior forecast and reconstruct the full thermo-chemical state.

2 Methodology

The proposed framework combines probabilistic reduced-order modelling, sparse sensing, and sequential data assimilation, in order to construct an uncertainty-aware digital twin capable of integrating numerical and experimental information within a unified Bayesian framework. First, the high-dimensional system dynamics are projected onto a low-dimensional latent space, where the temporal evolution of the reduced state is forecast using Gaussian Process Regression (GPR). Sparse measurements are then used to reconstruct the latent state, and an ensemble Kalman Filter (EnKF) is employed to correct the forecast by accounting for both model and observation uncertainties.

2.1 Reduced-order decomposition

To obtain a low-dimensional representation of the combustion dynamics, the original spatio-temporal dataset is arranged into the snapshot matrix

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_t}] \in \mathbb{R}^{N_p \times N_t}, \tag{1}$$

where each column $\mathbf{x}_t$ corresponds to a spatial snapshot acquired at time t. The dataset therefore contains N_t snapshots, each composed of N_p spatial features. Prior to the decomposition, the temporal mean is removed and standard-deviation scaling is applied to the data, leading to the preprocessed snapshot matrix $\widetilde{\mathbf{X}}$.

The objective of the reduced-order decomposition is to approximate the high-dimensional field using a limited number of spatial modes and their associated temporal coefficients. Since these temporal coefficients are subsequently regressed, it is important to obtain latent trajectories with interpretable temporal and frequency content. For this reason, the multiscale Proper Orthogonal Decomposition (mPOD) proposed by Mendez *et al.* [11] is adopted.

In contrast with classical POD, which ranks modes according to the global energetic content of the dataset, mPOD introduces a temporal scale separation by filtering the

temporal correlation matrix into given frequency bands before constructing the final modal basis. This reduces the mixing of structures evolving at different characteristic time scales and provides coefficients that are more suitable for frequency-informed temporal modelling. The temporal correlation matrix is defined as $\mathbf{K} = \widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}} \in \mathbb{R}^{N_t \times N_t}$. Following the mPOD formulation of Mendez *et al.* [11], this matrix is decomposed into N_b band-limited contributions through a set of filtering operators,

$$\mathcal{F}^{(b)}\left(\mathbf{K}\right), \qquad b = 1, \ldots, N_b, \tag{2}$$

where $\mathcal{F}^{(b)}$ denotes the filtering operator associated with the frequency band b. Each filtered correlation matrix is then independently decomposed as

$$\mathbf{K}^{(b)} \mathbf{\Psi}^{(b)} = \mathbf{\Psi}^{(b)} \mathbf{\Lambda}^{(b)}, \tag{3}$$

where $\mathbf{\Psi}^{(b)}$ contains the eigenvectors associated with the b frequency band and $\mathbf{\Lambda}^{(b)}$ is the diagonal matrix of the corresponding eigenvalues, which are proportional to the energy captured by the associated modes. The temporal bases obtained from all frequency bands are subsequently combined, ranked according to their modal energy and orthogonalized to construct a global temporal basis $\mathbf{\Psi}$. The associated spatial basis is then obtained by projecting the preprocessed snapshots matrix onto this temporal basis

$$\mathbf{\Phi}_r = \widetilde{\mathbf{X}} \mathbf{\Psi}_r \mathbf{\Lambda}_r^{-1/2}, \tag{4}$$

where $\mathbf{\Phi}_r \in \mathbb{R}^{N_p \times r}$ contains the retained spatial modes The temporal coefficients are computed by projecting the preprocessed snapshots onto the retained spatial basis,

$$\mathbf{A}_r = \mathbf{\Phi}_r^T \widetilde{\mathbf{X}}. \tag{5}$$

where $\mathbf{A}_r \in \mathbb{R}^{r \times N_t}$ contains the temporal evolution of the retained modes. Equivalently, using the eigendecomposition of the temporal correlation matrix, the temporal coefficients can be expressed as

$$\mathbf{A}_r = \mathbf{\Lambda}_r^{1/2} \mathbf{\Psi}_r^T. \tag{6}$$

Finally, the rank-r reduced-order representations of the dataset is written as

$$\widetilde{\mathbf{X}} \approx \mathbf{\Phi}_r \mathbf{A}_r, \tag{7}$$

or equivalently, for each snapshot at time t_j,

$$\widetilde{\mathbf{x}}(t_j) \approx \sum_{i=1}^{r} a_i(t_j) \boldsymbol{\phi}_i. \tag{8}$$

Here, $\mathbf{\Phi}_r = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_r]$ contains the retained spatial modes, $a_i(t_j) = (\mathbf{A}_r)_{i,j}$ is the temporal coefficient of mode i at time t_j, and r is the number of modes retained in the reduced representation. The temporal coefficients $a_i(t)$ therefore define the latent state subsequently forecast by the probabilistic regression model.

2.2 Probabilistic latent-state forecasting

The temporal evolution of the retained coefficients $a_i(t)$ is modeled using Gaussian Process Regression (GPR) [14]. An independent GPR is employed for each coefficient, providing a predictive posterior Gaussian distribution. For each mode i, the one-step reduced dynamics is formulated as

$$a_i(t_{j+1}) = f_i\left(\mathbf{a}(t_j), t_j\right), \tag{9}$$

where $\mathbf{a}(t_j) \in \mathbb{R}^r$ denotes the reduced state at time t_j, and $a_i(t_{j+1})$ is the one-step-ahead value of the ith reduced coefficient. A separate latent function f_i is defined for each retained coefficient and modelled as a Gaussian process,

$$f_i(\cdot) \sim \mathcal{GP}\left(m_i(\cdot), k_i(\cdot,\cdot)\right), \tag{10}$$

where m_i and k_i denote the prior mean and covariance function, respectively. In this work, a constant mean function $m_i(\cdot) = c_i$ is employed, with c_i learnt from the training data together with the kernel parameters.

For each retained modal coefficient, the Gaussian Process training dataset is defined as

$$\mathcal{D}_i = \{X_i, y_i\}, \quad X_i = \{[a_i(t_j),\, t_j]\}_{j=1}^{N_t-1}, \quad y_i = \{a_i(t_{j+1})\}_{j=1}^{N_t-1}. \tag{11}$$

Given a new input $X_{i,j}^* = [a_i(t_j^*),\, t_j^*]$, the i-th latent coefficient at the next time step $a_i(t_{j+1}^*)$ is modelled by the predictive distribution

$$a_i(t_{j+1}^*) \mid \mathcal{D}_i \sim \mathcal{N}\left(\mu_{i,j}^*, \gamma_{i,j}^*\right) \tag{12}$$

where $\mu_{i,j}^*$ and $\gamma_{i,j}^*$ denote the predictive mean and variance. These quantities follow the standard GP conditioning relations [14],

$$\begin{aligned} \mu_{i,j}^* &= K_i(X_{i,*}, X_i)\left[K_i(X_i, X_i) + \sigma_{n,i}^2\mathbf{I}\right]^{-1} y_i, \\ \gamma_{i,j}^* &= K_i(X_{i,*}, X_{i,*}) - K_i(X_{i,*}, X_i)\left[K_i(X_i, X_i) + \sigma_{n,i}^2\mathbf{I}\right]^{-1} K_i(X_i, X_{i,*}). \end{aligned} \tag{13}$$

Here, $K(X_i, X_i)$ is the covariance matrix evaluated over the training inputs per mode i, $K_i(X_{i,*}, X_i)$ is the cross-covariance between the prediction and training inputs, and $\sigma_{n,i}^2$ is the observation-noise variance estimated during training. Since each retained coefficient is modeled independently, the scalar predictive distributions are collected into a probabilistic latent-state forecast. For a given prediction time t_p, this forecast is written as

$$\mathbf{a}_{\mathrm{GPR}}(t_p) \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathrm{GPR},p}, \boldsymbol{\Gamma}_{\mathrm{GPR},p}\right) \tag{14}$$

with

$$\boldsymbol{\mu}_{\mathrm{GPR},p} = \begin{bmatrix}\mu_{1,p} & \cdots & \mu_{r,p}\end{bmatrix}^T, \qquad \boldsymbol{\Gamma}_{\mathrm{GPR},p} = \mathrm{diag}\left(\gamma_{1,p}, \ldots, \gamma_{r,p}\right). \tag{15}$$

The covariance function combines a state-dependent contribution with a frequency-informed temporal component,

$$k_i(\mathbf{z}, \mathbf{z}') = k_{\text{state},i}(\mathbf{a}, \mathbf{a}') + k_{\text{time},i}(t, t') + k_{\text{c},i}, \tag{16}$$

where a Matérn kernel is used for the reduced-state dependence, while different frequency-informed kernels are investigated for the temporal component.

As the reduced coefficients exhibit oscillatory dynamics, different frequency-informed kernels were investigated, including the Cosine, Periodic and Spectral Mixture kernels, to evaluate their capability to represent oscillatory combustion dynamics. The Cosine kernel represents a single dominant oscillatory component, the Periodic kernel imposes a repeated temporal structure, and the Spectral Mixture kernel provides a more flexible representation of multi-frequency dynamics. The periodic covariance follows the standard GP construction described by Rasmussen and Williams [14], whereas the Spectral Mixture formulation follows Wilson and Adams [24].

Denoting by $\tau = t - t'$ the temporal separation between two inputs, the three temporal kernels are defined as

$$\begin{aligned} k_{\cos}(\tau) &= \sigma_f^2 \cos\left(2\pi f_0 \tau\right), \\ k_{\text{per}}(\tau) &= \sigma_f^2 \exp\left[-\frac{2\sin^2\left(\pi\tau/T\right)}{\ell^2}\right], \\ k_{\text{SM}}(\tau) &= \sum_{q=1}^{Q} w_q \exp\left(-2\pi^2 v_q \tau^2\right) \cos\left(2\pi\mu_q \tau\right). \end{aligned} \tag{17}$$

Here, σ_f^2 controls the signal variance, f_0 denotes the dominant frequency of the Cosine kernel, T is the period of the Periodic kernel, and ℓ controls the correlation length within each period. For the Spectral Mixture kernel, Q is the number of mixture components, while w_q, μ_q, and v_q denote the weight, mean frequency, and spectral variance of the q-th component, respectively.

The kernels hyperparameters are optimized independently for each modal GPR. For Cosine and Periodic kernels, the temporal period is initialized from the dominant frequency f_0, identified by mPOD such that $T_{0,i} = 1/f_{0,i}$, and is subsequently optimized within a frequency interval of ± 50 Hz around f_0. For the Spectral Mixture kernel, the number of components is fixed to $Q = 3$, their mean frequencies initialized at $0.75f_0$, f_0 and $1.25f_0$, while w_q, μ_q, and v_q and the remaining kernel hyperparameters are then learned during training.

2.3 Data assimilation

After the GPR forecast of the mPOD coefficients has been obtained, data assimilation is employed to correct the predicted latent trajectories using newly available measurements. The correction is performed directly in the reduced space, so that only the temporal coefficients are updated before reconstructing the physical fields. The following subsections describe how sparse measurements are converted into latent observations, how their uncertainty is estimated, and how this information will be incorporated into the correction.

2.3.1 Sparse sensing and latent-state reconstruction

Using the reduced basis introduced in the previous section, sparse sensing is employed to infer the latent state of the system from a limited number of spatial measurements of the physical state [15]. The objective is to reconstruct mPOD temporal coefficients a from sparse observations, thereby providing reduced-space state estimates for the subsequent data assimilation step.

At each time instant t_j, the vector of sparse measurements is defined as:

$$\mathbf{y}(t_j) = \mathbf{C}\widetilde{\mathbf{x}}(t_j) \tag{18}$$

where $\mathbf{C} \in \mathbb{R}^{p \times N_p}$ is the measurement matrix selecting the p sensor locations. Introducing the rank-r reduced representation, the sensing problem becomes

$$\mathbf{y}(t_j) \approx \mathbf{\Theta}\mathbf{a}(t_j), \qquad \mathbf{\Theta} = \mathbf{C}\mathbf{\Phi}_r \tag{19}$$

where $\mathbf{\Theta} \in \mathbb{R}^{p \times r}$ is the reduced sensing matrix, and $\mathbf{\Phi}_r$ is the matrix of the retained spatial modes. The reconstruction of the latent coefficients therefore corresponds to the inverse problem of estimating $\mathbf{a}(t_j)$ from the sparse measurement vector $\mathbf{y}(t_j)$. Three reconstruction strategies, providing both a latent estimate and an associated uncertainty measure, are considered in this work: weighted least squares, regularized least squares, and Bayesian ridge regression.

Weighted least squares

The weighted least squares (WLS) formulation accounts explicitly for the uncertainty associated with the sensor measurements. The measurement-noise covariance is defined as

$$\mathbf{\Gamma}_y = \mathrm{diag}\left(\widetilde{\eta}_1^2, \ldots, \widetilde{\eta}_p^2\right), \tag{20}$$

where $\widetilde{\eta}_s$ denotes the scaled uncertainty associated with the s-th selected sensor, with $s = 1, \ldots, p$. The uncertainties were estimated directly from the experimental data using a clustering-based approach. First, the original snapshots were projected onto a Principal Component Analysis (PCA) representation [25] retaining $r = 6$ components. Then, the snapshots were grouped into 3 clusters using k-means [26] in the PCA space, identifying groups of similar dynamical states. For a given cluster c, the effective uncertainty at the s-th selected sensor, $\eta_s^{(c)}$, was computed from the original unscaled data as:

$$\eta_s^{(c)} = \sqrt{\frac{1}{N_c} \sum_{j \in \mathcal{C}_c} \left[X_{i_s,j} - \overline{X}_{i_s}^{(c)}\right]^2}, \qquad \overline{X}_{i_s}^{(c)} = \frac{1}{N_c} \sum_{j \in \mathcal{C}_c} X_{i_s,j}, \tag{21}$$

where $\mathcal{C}_c$ denotes the set of snapshots assigned to cluster c, N_c is the number of snapshots in that cluster, $X_{i_s,j}$ denotes the chemiluminescence intensity at pixel i_s, corresponding to the s-th selected sensor, in snapshot j. The quantity $\overline{X}_{i_s}^{(c)}$ is the mean intensity at that pixel across the snapshots belonging to cluster c. Thus, $\eta_s^{(c)}$ quantifies the local intensity variability at the selected sensor location within cluster c. For the subsequent analysis, the cluster exhibiting the lowest intra-cluster variability, namely,

$c = 1$, was selected to define the sensor uncertainties. Since the sparse measurements used in the WLS formulation are expressed in scaled coordinates, the uncertainties estimated from the original data were scaled using the same scaling factors as those applied to the measurement data. The resulting scaled uncertainties, denoted by $\widetilde{\eta}_s$, are therefore consistent with the scaled measurement vector $\mathbf{y}(t_j)$ (see Eq. 18) and were used to construct the diagonal covariance matrix $\mathbf{\Gamma}_y$ in Eq. 20.

Using the measurement model in Eq. 19, the latent coefficients are estimated by minimizing the weighted measurement residual,

$$\widehat{\mathbf{a}}_{\mathrm{WLS}}(t_j) = \arg\min_{\mathbf{a}} \left[\mathbf{y}(t_j) - \mathbf{\Theta}\mathbf{a}\right]^T \mathbf{\Gamma}_y^{-1} \left[\mathbf{y}(t_j) - \mathbf{\Theta}\mathbf{a}\right]. \tag{22}$$

Assuming that $\mathbf{\Theta}$ has full column rank, which requires $p \geq r$, Eq. 22 admits the unique solution:

$$\widehat{\mathbf{a}}_{\mathrm{WLS}}(t_j) = \left(\mathbf{\Theta}^T \mathbf{\Gamma}_y^{-1} \mathbf{\Theta}\right)^{-1} \mathbf{\Theta}^T \mathbf{\Gamma}_y^{-1} \mathbf{y}(t_j). \tag{23}$$

For $p > r$ the problem is overdetermined, whereas for $p = r$ a unique solution exists only if $\mathbf{\Theta}$ is nonsingular. For $p < r$, the problem is undetermined and the inverse in Eq. 23 is not defined, requiring additional regularization through the incorporation of prior information.

Under the linear-Gaussian assumption, the corresponding covariance of the estimated coefficients is

$$\mathbf{\Gamma}_{a,\mathrm{WLS}} = \left(\mathbf{\Theta}^T \mathbf{\Gamma}_y^{-1} \mathbf{\Theta}\right)^{-1}. \tag{24}$$

Regularized least squares

The regularized least squares (RLS) formulation extends the WLS inverse problem in Eq. 22 by introducing prior information on the expected magnitude of the reduced coefficients [27, 28]. Specifically, a Gaussian prior is assumed for the latent coefficients,

$$\mathbf{a} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{S}^2\right), \tag{25}$$

where $\mathbf{S} = \mathrm{diag}(s_1, \ldots, s_r)$ is a diagonal scale matrix defining the characteristic magnitude of each reduced coefficient. Using the same measurement covariance $\mathbf{\Gamma}_y$ introduced in Eq. 20, the corresponding maximum-a-posteriori estimate is obtained by solving

$$\widehat{\mathbf{a}}_{\mathrm{RLS}}(t_j) = \arg\min_{\mathbf{a}} \left[\left(\mathbf{y}(t_j) - \mathbf{\Theta}\mathbf{a}\right)^T \mathbf{\Gamma}_y^{-1} \left(\mathbf{y}(t_j) - \mathbf{\Theta}\mathbf{a}\right) + \mathbf{a}^T \mathbf{S}^{-2} \mathbf{a}\right]. \tag{26}$$

Compared with the WLS formulation in Eq. 22, the additional term penalizes solutions that are inconsistent with the prescribed coefficient scales. The resulting analytical estimate is

$$\widehat{\mathbf{a}}_{\mathrm{RLS}}(t_j) = \mathbf{A}_{\mathrm{RLS}} \mathbf{y}(t_j). \tag{27}$$

with

$$\mathbf{A}_{\mathrm{RLS}} = \left(\mathbf{S}^{-2} + \mathbf{\Theta}^T \mathbf{\Gamma}_y^{-1} \mathbf{\Theta}\right)^{-1} \mathbf{\Theta}^T \mathbf{\Gamma}_y^{-1}. \tag{28}$$

The corresponding posterior covariance in the reduced space is

$$\boldsymbol{\Gamma}_{a,\mathrm{RLS}} = \left(\mathbf{S}^{-2} + \boldsymbol{\Theta}^T \boldsymbol{\Gamma}_y^{-1} \boldsymbol{\Theta}\right)^{-1}. \tag{29}$$

Unlike WLS, the regularized formulation does not require $\boldsymbol{\Theta}$ to have full column rank. For $p > r$, regularization improves robustness when the overdetermined system is ill-conditioned; for $p = r$, it stabilizes square sensing matrices that may be poorly conditioned; and for $p < r$, the prior term converts an undetermined problem into a well-posed one, allowing a stable estimate of the latent coefficient.

Bayesian ridge regression

An alternative reconstruction strategy is obtained with Bayesian Ridge Regression (BRR) [29]. Rather than directly inverting the sensing model in Eq. 19, BRR learns the mapping between sparse measurements and each temporal coefficient from the training data:

$$\mathbf{Y}_{\mathrm{tr}} = [\mathbf{y}(t_1), \ldots, \mathbf{y}(t_{N_{\mathrm{tr}}})] \in \mathbb{R}^{p \times N_{\mathrm{tr}}}, \tag{30}$$

where $\mathbf{y}$ denotes the sparse measurements available at each training snapshot. For each mode i, the regression model is

$$\mathbf{a}_{i,\mathrm{tr}} = \mathbf{Y}_{\mathrm{tr}}^T \mathbf{w}_i + \boldsymbol{\varepsilon}_i, \tag{31}$$

where $\mathbf{w}_i \in \mathbb{R}^p$ contains the regression weights. A Gaussian prior is assigned to these weights,

$$p(\mathbf{w}_i) = \mathcal{N}\left(\mathbf{0}, \lambda_i^{-1}\mathbf{I}\right), \tag{32}$$

while the residuals in Eq. 31 are assumed Gaussian with precision α_i. The resulting posterior covariance and mean of the regression weights are

$$\boldsymbol{\Sigma}_{w_i} = \left(\lambda_i \mathbf{I} + \alpha_i \mathbf{Y}_{\mathrm{tr}} \mathbf{Y}_{\mathrm{tr}}^T\right)^{-1}, \tag{33}$$

$$\overline{\mathbf{w}}_i = \alpha_i \boldsymbol{\Sigma}_{w_i} \mathbf{Y}_{\mathrm{tr}} \mathbf{a}_{i,\mathrm{tr}}, \tag{34}$$

where α_i and λ_i are inferred during model fitting. Once the mapping has been trained, a new sparse measurement is projected into the latent space as

$$\widehat{a}_{\mathrm{BR},i}(t_j) = \overline{\mathbf{w}}_i^T \mathbf{y}(t_j). \tag{35}$$

The Bayesian formulation also provides a predictive uncertainty for each reconstructed coefficient. For a new sparse measurement $\mathbf{y}(t_j)$, the predictive distribution is

$$a_{\mathrm{BR},i}(t_j) \mid \mathbf{y}(t_j), \mathcal{D}_i \sim \mathcal{N}\left(\widehat{a}_{\mathrm{BR},i}(t_j), \gamma_{\mathrm{BR},i}(t_j)\right), \tag{36}$$

with

$$\gamma_{\mathrm{BR},i}(t_j) = \alpha_i^{-1} + \mathbf{y}(t_j)^T \boldsymbol{\Sigma}_{w_i} \mathbf{y}(t_j). \tag{37}$$

Collecting the regression weights for all r modes into $\mathbf{W}_{\mathrm{BR}} \in \mathbb{R}^{r \times p}$ gives

$$\widehat{\mathbf{a}}_{\mathrm{BR}}(t_j) = \mathbf{W}_{\mathrm{BR}} \mathbf{y}(t_j). \tag{38}$$

Since each modal coefficient is reconstructed independently, the corresponding latent-state covariance is assembled as

$$\mathbf{\Gamma}_{\mathrm{BR},j} = \mathrm{diag}\left(\gamma_{\mathrm{BR},1}(t_j), \ldots, \gamma_{\mathrm{BR},r}(t_j)\right). \tag{39}$$

Consequently, WLS and RLS reconstruct the coefficients by solving the sparse inverse problem defined by $\mathbf{\Theta}$, whereas BRR learns an inverse measurement-to-latent mapping from paired training data. Although the three strategies produce the same reduced-state quantity $\widehat{\mathbf{a}}$, they represent distinct reconstruction approaches and are therefore compared empirically in this work.

2.3.2 Uncertainty quantification

The uncertainty associated with both the latent-state prediction and the sparse-sensing reconstruction enables the formulation of a probabilistic data-assimilation framework.

At each time instant t_j, the GPR model provides the predictive covariance $\mathbf{\Gamma}_{\mathrm{GPR},j}$ associated with the forecasted latent state. Each sparse reconstruction method also provides an estimate of the uncertainty in the reconstructed latent state. The corresponding baseline covariance is denoted as $\mathbf{\Gamma}_{\mathrm{SPS}}$, and its specific expression depends on the reconstruction method adopted, such as Eqs. 24, 29 and 39.

To introduce a time-dependent measure of reconstruction confidence, the baseline covariance is inflated by a scalar factor λ_j,

$$\mathbf{\Gamma}_{\mathrm{SPS},j} = \lambda_j \mathbf{\Gamma}_{\mathrm{SPS}}. \tag{40}$$

The inflation factor is computed from the consistency between the available measurements and their reconstruction from the estimated latent coefficients. The measurement residual is

$$\mathbf{e}_j = \mathbf{y}(t_j) - \widehat{\mathbf{y}}(t_j) = \mathbf{y}(t_j) - \mathbf{\Theta}\widehat{\mathbf{a}}_{\mathrm{SPS}}(t_j) \tag{41}$$

A normalized residual statistic is then defined as

$$\chi_j^2 = \frac{1}{p-r}\mathbf{e}_j^T \mathbf{\Gamma}_y^{-1} \mathbf{e}_j = \frac{1}{p-r}\sum_{s=1}^{p}\left(\frac{e_{j,s}}{\eta_s}\right)^2, \qquad p > r. \tag{42}$$

The covariance inflation factor is chosen as

$$\lambda_j = \max\left(1, \chi_j^2\right) \tag{43}$$

Thus, snapshots exhibiting larger weighted reconstruction residuals are assigned a larger reconstruction uncertainty, whereas snapshots consistent with the sparse measurement model retain the baseline covariance. The resulting sparse latent-state estimate is represented probabilistically as

$$\mathbf{a}_{\mathrm{SPS}}(t_j) \sim \mathcal{N}\left(\widehat{\mathbf{a}}_{\mathrm{SPS}}(t_j), \mathbf{\Gamma}_{\mathrm{SPS},j}\right). \tag{44}$$

The covariance inflation introduced above is used as an adaptive uncertainty model and is not intended to represent an exact posterior covariance of the latent-state reconstruction.

2.3.3 Ensemble Kalman filter

The uncertainty estimates introduced in the previous section enable the GPR-based forecast and the sparse latent-state reconstruction, to be combined through an Ensemble Kalman Filter (EnKF) [30]. Data assimilation is performed directly in the reduced space, where the stochastic GPR forecast propagates an ensemble of latent states and the sparse sensing procedure provides probabilistic latent observations. In the following, the superscripts f and a denote forecast and analysis states, respectively.

At each assimilation time t_j, the forecast ensemble is written as

$$\mathbf{A}_j^f = \left[\mathbf{a}_j^{f,(1)}, \ldots, \mathbf{a}_j^{f,(N_e)}\right] \in \mathbb{R}^{r \times N_e}, \tag{45}$$

where N_e is the ensemble size and each $\mathbf{a}_j^{f,(n)} \in \mathbb{R}^r$ represents one realization of the stochastic GPR forecast. The ensemble mean is

$$\overline{\mathbf{a}}_j^f = \frac{1}{N_e} \sum_{n=1}^{N_e} \mathbf{a}_j^{f,(n)}, \tag{46}$$

and the corresponding sample covariance is estimated from the ensemble as

$$\mathbf{\Gamma}_{\mathrm{GPR},j}^{\mathrm{ens}} = \frac{1}{N_e - 1} \sum_{n=1}^{N_e} \left(\mathbf{a}_j^{f,(n)} - \overline{\mathbf{a}}_j^f\right) \left(\mathbf{a}_j^{f,(n)} - \overline{\mathbf{a}}_j^f\right)^T. \tag{47}$$

While $\mathbf{\Gamma}_{\mathrm{GPR},j}$ denotes the predictive covariance provided directly by the independent GPR models, as defined in Eq. 15, $\mathbf{\Gamma}_{\mathrm{GPR},j}^{\mathrm{ens}}$ represents the sample covariance of the recursively propagated forecast ensemble. Although the GPR models are trained independently, the recursive stochastic propagation can introduce cross-correlations between the retained coefficients, which are captured by $\mathbf{\Gamma}_{\mathrm{GPR},j}^{\mathrm{ens}}$.

As shown in Eq. 44, the sparse sensing procedure provides latent observations characterized by its mean $\widehat{\mathbf{a}}_{\mathrm{SPS}}(t_j)$ and covariance $\mathbf{\Gamma}_{\mathrm{SPS},j}$. Since the forecast and sparse observations are expressed in the same reduced coordinate system, the observation operator is the identity and the corresponding Kalman gain becomes

$$\mathbf{K}_{\mathrm{DA},j} = \mathbf{\Gamma}_{\mathrm{GPR},j}^{\mathrm{ens}} \left(\mathbf{\Gamma}_{\mathrm{GPR},j}^{\mathrm{ens}} + \mathbf{\Gamma}_{\mathrm{SPS},j}\right)^{-1}. \tag{48}$$

A stochastic, perturbed observation formulation of the EnKF is then employed. For each ensemble member, an independent realization of the sparse latent observation is drawn as

$$\mathbf{a}_{\mathrm{SPS},j}^{(n)} = \widehat{\mathbf{a}}_{\mathrm{SPS}}(t_j) + \boldsymbol{\epsilon}_j^{(n)}, \qquad \boldsymbol{\epsilon}_j^{(n)} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Gamma}_{\mathrm{SPS},j}\right), \tag{49}$$

and each forecast ensemble member is updated according to

$$\mathbf{a}_j^{a,(n)} = \mathbf{a}_j^{f,(n)} + \mathbf{K}_{\mathrm{DA},j} \left[\mathbf{a}_{\mathrm{SPS},j}^{(n)} - \mathbf{a}_j^{f,(n)}\right]. \tag{50}$$

The perturbed observations incorporate the sparse reconstruction uncertainty into the ensemble update. The corrected latent state is represented by the analysis ensemble mean,

$$\overline{\mathbf{a}}_j^a = \frac{1}{N_e} \sum_{n=1}^{N_e} \mathbf{a}_j^{a,(n)}. \tag{51}$$

Following each assimilation step, the analysis ensemble is propagated forward using the stochastic GPR model until the next sparse observation becomes available. Finally, the corrected physical state is reconstructed from the reduced basis as

$$\widehat{\mathbf{x}}_{\mathrm{DA}}(t_j) = \mathbf{\Phi}_r \overline{\mathbf{a}}_j^a. \tag{52}$$

2.4 Performance metrics

The prediction and reconstruction errors are quantified using the normalized root mean squared error (NRMSE). For a reference signal y and its prediction $\widehat{y}$, the RMSE is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (\widehat{y}_k - y_k)^2}. \tag{53}$$

To enable comparison among quantities with different amplitudes, the error is normalized by the range of the reference signal,

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\max(\boldsymbol{y}) - \min(\boldsymbol{y})}. \tag{54}$$

In cases where a single metric is required for the complete reduced state, the modal errors are combined using their relative mPOD energy,

$$w_i = \frac{\lambda_i}{\sum_{j=1}^{r} \lambda_j}, \qquad \mathrm{NRMSE}_w = \sqrt{\sum_{i=1}^{r} w_i \, \mathrm{NRMSE}_i^2}. \tag{55}$$

where λ_i denotes the energy of mode i and $\sum_{i=1}^{r} w_i = 1$.

3 Datasets

The two datasets presented below provide distinct test cases, each offering challenges to the proposed framework. The experimental dataset contains noisy images of a periodically forced flame in a combustion chamber and reflects practical sensing constraints, whereas the numerical dataset involves a fully resolved reacting flow with a more complex multi-frequency dynamics, allowing the assimilation of *numerical* measurements with a controlled level of noise and the recovery of the complete thermo-chemical state.

3.1 Experimental dataset

The first dataset consists of time-resolved CH* chemiluminescence images obtained inside the BIMER combustor, shown in Figure 1, which has been designed to study a swirled two-stage injection strategy [31]. Air enters a cylindrical plenum and flows

into a rectangular combustion chamber (500 × 150 × 150 mm^3) with three water-cooled walls (entrance, top, bottom) and two lateral quartz windows for optical access. The injector has a central pilot stage, which produces a hollow cone of fuel droplets (30° spray half-angle) and imposes a geometric swirl number of 0.6, and a multipoint stage that injects fuel through ten 0.3 mm-diameter holes at the second swirler vanes (swirl number of 1), creating jets in cross-flow to enhance mixing. The fuel is liquid dodecane, acting as a surrogate for kerosene.

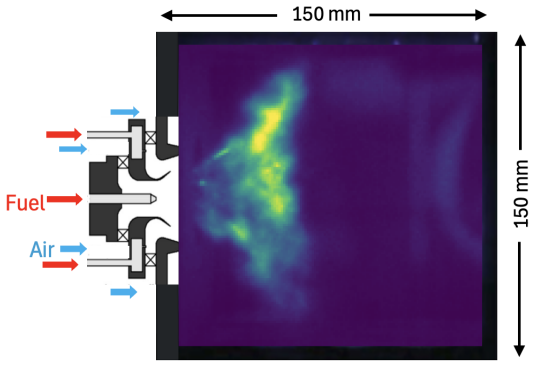


Fig. 1: View of the BIMER set-up.

The experimental sequence used in this study captures the time evolution of a swirled flame subjected to a 100Hz modulation of the inlet air, while the dodecane is injected only through the pilot stage. The dataset comprises 5458 snapshots corresponding to approximately 1.09 s of flame evolution. Each image was spatially resized from 512×512 to 64×64 pixels to reduce the spatial dimensionality and computational cost of the reduced-order analysis, while retaining the dominant flame structures relevant to the latent dynamics. The forcing frequency induces a strongly oscillatory flame response, accompanied by experimental noise and additional dynamical components. This dataset therefore provides an experimental benchmark for evaluating the capability of the ROM to forecast noisy combustion dynamics. The BIMER dataset has previously been employed to assess a stochastic GPR-based ROM for the prediction of noisy combustion dynamics [23].

3.2 Numerical dataset

This dataset corresponds to high-fidelity simulations of an axisymmetric, time-dependent, non-premixed laminar coflow methane flame, which was previously targeted for reduced-order modelling of reacting flows by Corrochano *et al.* [32]. The simulations were performed with the detailed chemical mechanism GRI-Mech 3.0 [33]. In the present analysis, temperature and species fields are retained in order to construct the reduced-order representation. The considered case is subjected to a multi-frequency modulation of the fuel inlet, thereby producing a more complex

temporal response than a single-frequency oscillation. The dataset contains 2000 snapshots, with a uniform time step of 0.5 ms and a total simulated interval of 1 s. The numerical configuration is based on the laminar reacting-flow database described by Corrochano *et al.* [34].

4 Results and discussion

4.1 Experimental case

Prior to the GPR training, mPOD was applied to the dataset to introduce a frequency-band separation. The reduced representation was truncated at the first six modes, retaining approximately 80% of the total variance. The corresponding spatial modes and temporal coefficients are shown in Figure 2. The six retained coefficients exhibit distinct dominant oscillatory scales. Since the BIMER combustor is forced at 100Hz, coefficients a_1 and a_3 are mainly associated with the fundamental forcing frequency. The reduced dynamics also contain higher-order harmonic contributions, which are represented by a_2 and a_4 at approximately 200Hz and by a_5 and a_6 at 300Hz.

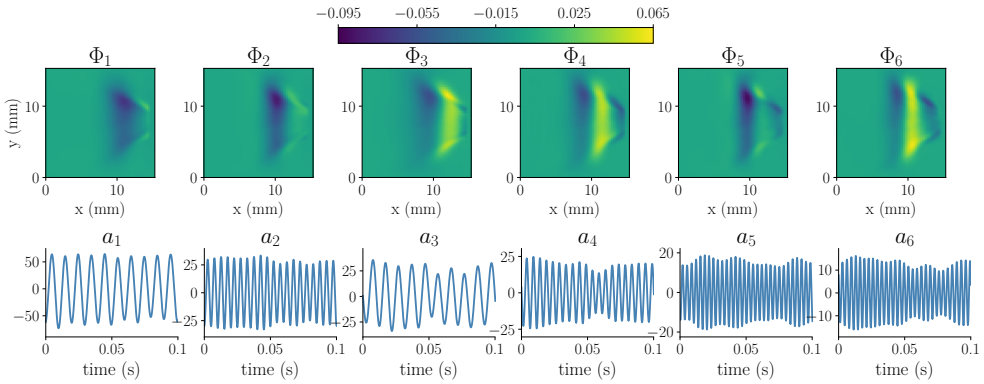
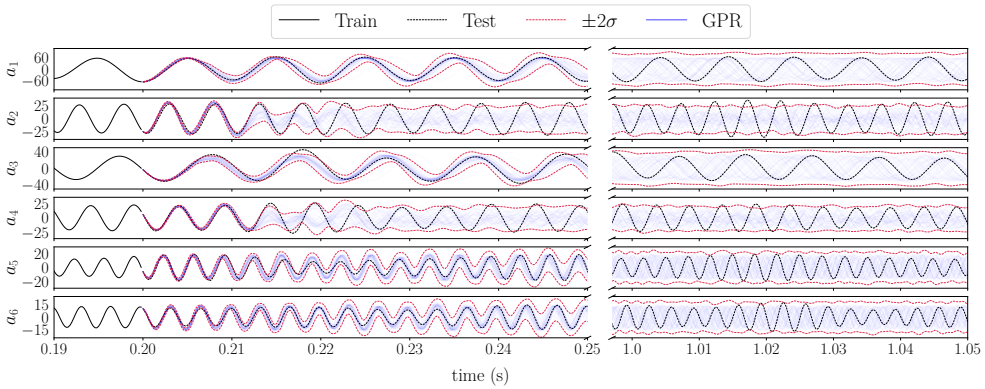


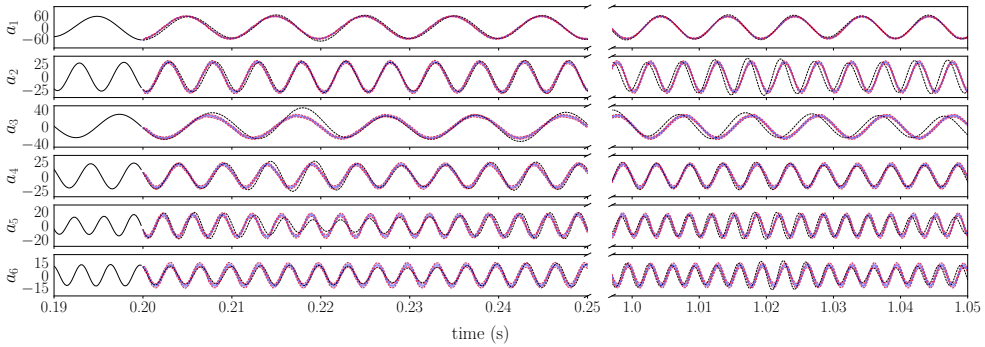
Fig. 2: First six mPOD modes retained for the BIMER dataset. The top row shows the spatial modes Φ_i, while the bottom row shows the corresponding temporal coefficients $a_i(t)$ over a representative 0.1 s interval.

4.1.1 GPR perfomance

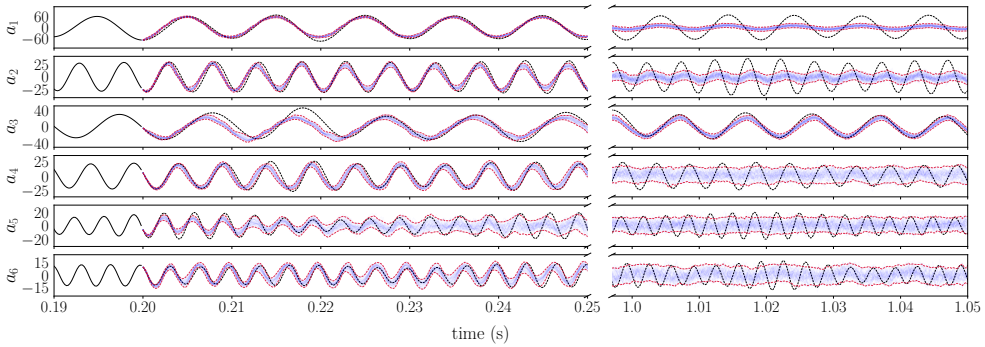
This spectral organization is particularly suitable for the subsequent GPR modelling, since the dominant frequency of each coefficient can be explicitly incorporated into the temporal covariance kernel. The comparison among the different covariance functions presented in Figure 3 shows that the kernel formulation has a significant impact on the forecast of the temporal coefficients. Over the prediction horizon, the Periodic kernel provides the most consistent agreement with the reference dynamics, particularly in preserving the dominant oscillation frequency and phase. In contrast, the Cosine kernel exhibits clear deviations for several coefficients, while Spectral Mixture kernel generally provides an intermediate performance.



(a) Cosine kernel



(b) Periodic kernel



(c) Spectral Mixture kernel

Fig. 3: Forecast of the first 6 temporal coefficients from 0.2 to 0.25 s and from 1.0 to 1.05 s, using different GPR kernel formulations. The black curves represent the true trajectories, the blue curves depict 100 trajectories predicted by GPR, and the red dashed lines indicate the GPR $\pm 2\sigma$ uncertainty.

These observations are supported by the NRMSE values defined in Section 2.4 and reported in Figure 4, which shows that the Periodic kernel consistently yields the lowest error across the retained coefficients. The advantage of the Periodic formulation can be attributed to its capacity to enforce the characteristic period of each mPOD coefficient, while retaining some flexibility through its correlation length. By comparison, the Cosine kernel imposes a more restrictive sinusoidal covariance structure, whereas the additional flexibility of the Spectral Mixture kernel does not provide a systematic benefit for the relatively narrow-band latent dynamics produced by mPOD. The findings suggest that the explicit incorporation of the dominant periodicity identified from the training data provides a suitable representation of the BIMER dynamics in the latent space.

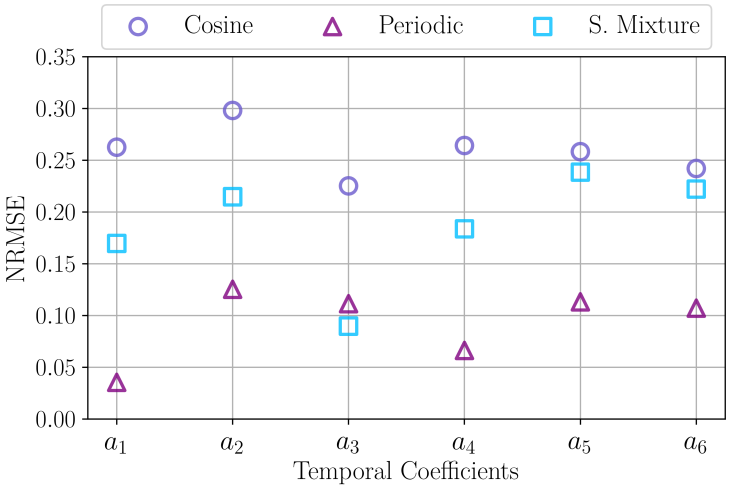


Fig. 4: NRMSE of the temporal-coefficient forecasts for different GPR kernels.

The results indicate that the forecast accuracy therefore strongly depends on how the kernel structure represents the temporal dynamics observed within the selected training interval.

4.1.2 Sparse sensing

Three sensing configurations were considered to assess the influence of spatial measurement availability on latent-state reconstruction:

(a) ***Window*** represents a limited field of view, as might occur in chemiluminescence imaging, where measurements are only available within a continuous $w \times w$ pixels region of the full system.

(b) ***QR + Random*** assumes access to the entire spatial domain. Informative sensor locations were first identified through QR-based selection using a number of sensors equal to the number of retained modes [15], while additional measurements were randomly distributed over the remaining domain.

(c) ***Window-constrained*** combines limited spatial accessibility with a finite number of measurements. Sensor locations were restricted to the accessible $w \times w$ region, allowing the effect of sensor sparsity and spatial constraints to be evaluated simultaneously.

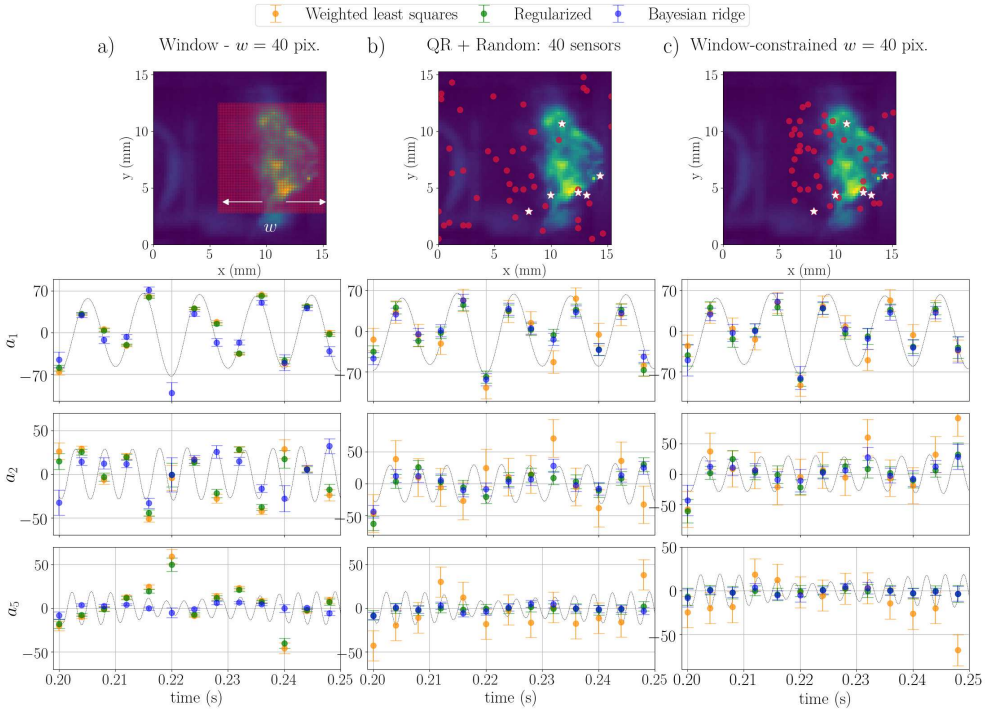
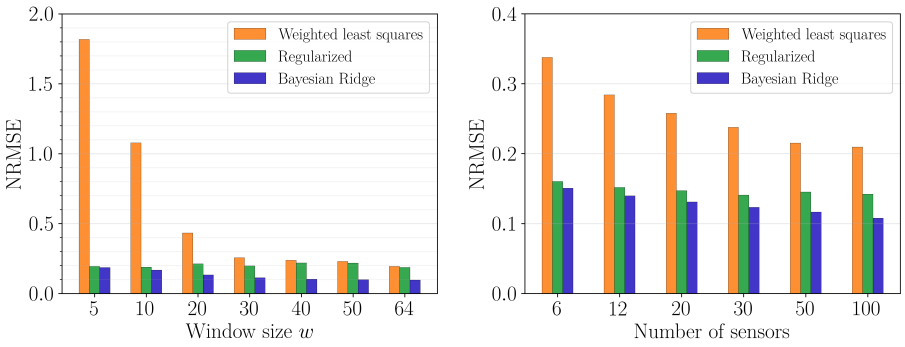


Fig. 5: Latent-state reconstruction and associated uncertainty for representative temporal coefficients under different sensing strategies: (a) Window, (b) QR + Random, and (c) Window-constrained. Both the red dots and the white stars represent sensor locations. Red dots correspond to random locations while white stars represent optimal positions according to the QR selection.

The three sensing strategies are shown in Figure 5 and they were implemented through the Pysensors2.0 package [28]. For each strategy, the available measurements were mapped into the latent space using the WLS, RLS and Bayesian ridge reconstruction approaches introduced in Section 2.3.1. The representative coefficients a_1, a_2, and a_5 are selected to illustrate the reconstruction of dynamics associated with the different

frequency bands identified by mPOD. The comparison therefore highlights the combined influence of spatial accessibility, sensor placement, and inverse reconstruction method on both the estimated latent state and its associated uncertainty.



(a) Window-based sensing

(b) QR + Random sensing

Fig. 6: NRMSE of the latent-state reconstruction for different sparse sensing configuration and inverse methods: (a) window-based sensing, where w is defined as a value measured in pixels; and (b) QR-based placement complemented with random sensors.

The reconstruction accuracy depends strongly on both the sensing configuration and the inverse method as shown in Figure 6. WLS is particularly sensitive to restrictive sensing conditions. For the window-based configuration, its error increases sharply as the accessible region is reduced, whereas Bayesian ridge and regularized approaches remain considerably more stable. A similar trend is observed for the QR + Random configuration: increasing the number of available sensors reduces the reconstruction error, although the improvement is substantially more pronounced for WLS. Bayesian ridge consistently provides the lowest NRMSE among the investigated methods, while the difference between the reconstruction approaches decreases as more spatial information becomes available.

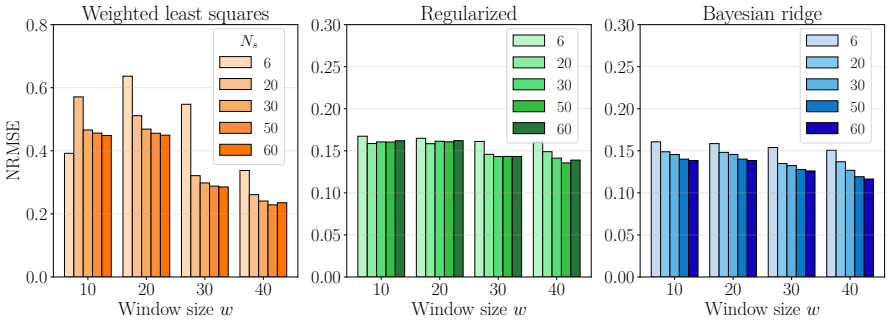


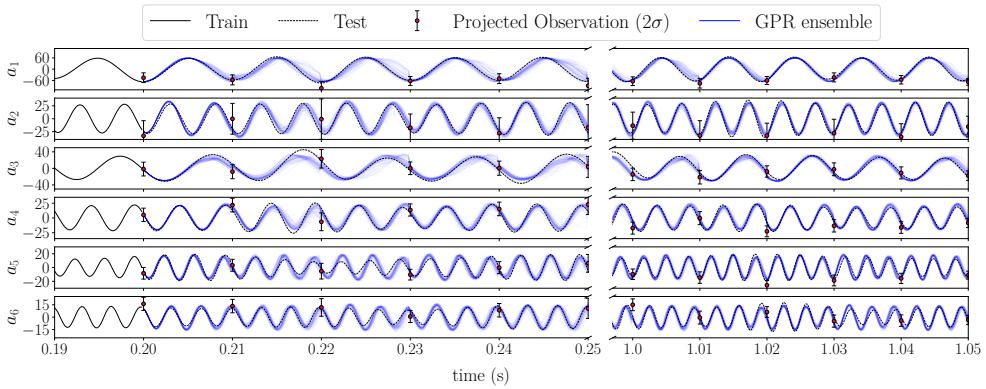
Fig. 7: NRMSE of the latent-state reconstruction of the window-constrained method as a function of window size, number of sensors, and reconstruction strategy, i.e., WLS, RLS, and Bayesian ridge.

The window-constrained results in Fig. 7 further show that increasing the number of sensors does not necessarily produce the same improvement for all reconstruction methods. For WLS, the reconstruction remains strongly dependent on the window size and exhibits a non-monotonic response to the number of measurements, indicating that the spatial distribution and conditioning of the reduced sensing matrix are as important as the number of sensors itself. In contrast, RLS is considerably less sensitive to these variations, while Bayesian Ridge shows a more systematic reduction in error as both the accessible region and the number of measurements increase. These results demonstrate the stabilizing effect of regularization when measurements are spatially restricted.

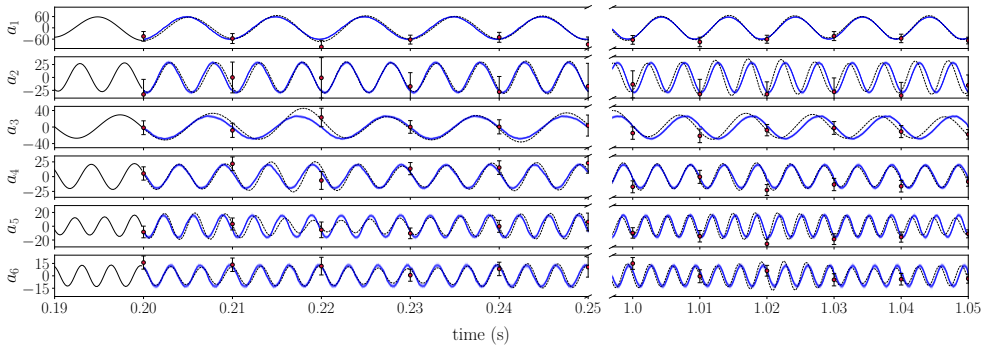
Based on the previous comparison, the window configuration with $w = 40$ pixel was selected for the subsequent data assimilation analysis. This configuration provides a suitable compromise between limited spatial accessibility and latent-state reconstruction accuracy, while remaining representative of the restricted field-of-view conditions considered for the BIMER measurements. The corresponding sparse latent-state estimates are therefore used as observations in the subsequent Kalman-filter-based correction stage.

4.1.3 Data assimilation

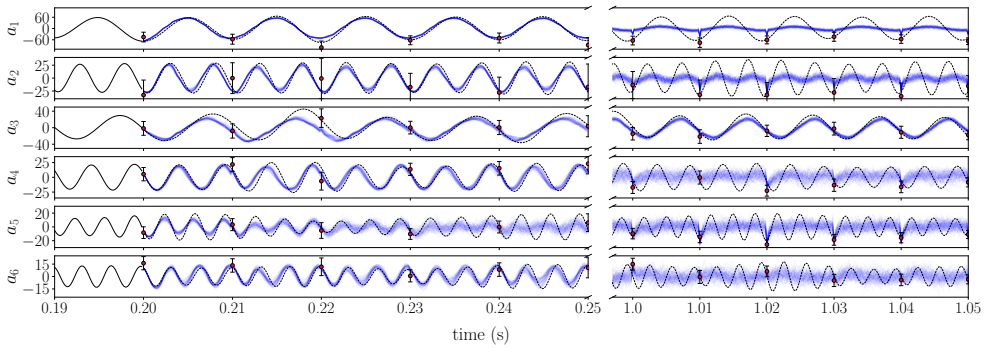
The sparse latent-state estimates obtained from the selected 40×40 pixel window were subsequently assimilated into the GPR forecast through the EnKF formulation introduced in Section 2.3.3. Figure 8 shows the corrected trajectories obtained for the Cosine, Periodic, and Spectral Mixture kernels using an assimilation interval of $50\Delta t = 0.01$ s. Between consecutive updates, the latent state evolves according to the GPR forecast, while the sparse sensing estimates periodically correct the predicted trajectories according to their uncertainty.



(a) Cosine kernel



(b) Periodic kernel



(c) Spectral Mixture kernel

Fig. 8: EnKF correction of the six temporal coefficients with the different GPR kernels. The red dots represent the observations assimilated every $50\Delta t = 0.01$ s with their uncertainty (black bars).

The effect of data assimilation depends strongly on the accuracy of the underlying GPR model and the quality of the sparse observations in the latent space. For the Periodic kernel, which already provides the most accurate baseline forecast, the EnKF introduces comparatively small improvements, such as a higher confidence (lower uncertainty) in the forecast. In contrast, the Cosine kernel benefits more substantially from assimilation, particularly for the temporal coefficients whose predicted trajectories exhibit amplitude or phase discrepancies relative to the reference. In this case, assimilation also prevents the forecast from diverging over extended time horizons.

The Spectral Mixture kernel exhibits a different behaviour. Although data assimilation reduces the uncertainty associated with the predictions, it does not improve their accuracy and can even slightly degrade the forecast. Before assimilation, the uncertainty bounds encompass the true trajectory, whereas the reduced uncertainty after assimilation makes the discrepancy between the predicted and true trajectories more apparent. This suggests that, in this case, the EnKF correction increases confidence in the forecast without a corresponding improvement in its accuracy, resulting in a somewhat overconfident prediction.

Figure 9 compares the NRMSE of the six temporal coefficients before and after data assimilation, while Table 1 reports the corresponding reconstruction errors in the physical fields. The influence of the assimilation frequency was evaluated by comparing steps of $10\Delta t$, $50\Delta t$ and $100\Delta t$. The results confirm a marked reduction in error for the Cosine kernel, whereas assimilation has a comparatively limited effect on the performance of the Periodic and Spectral Mixture kernels.

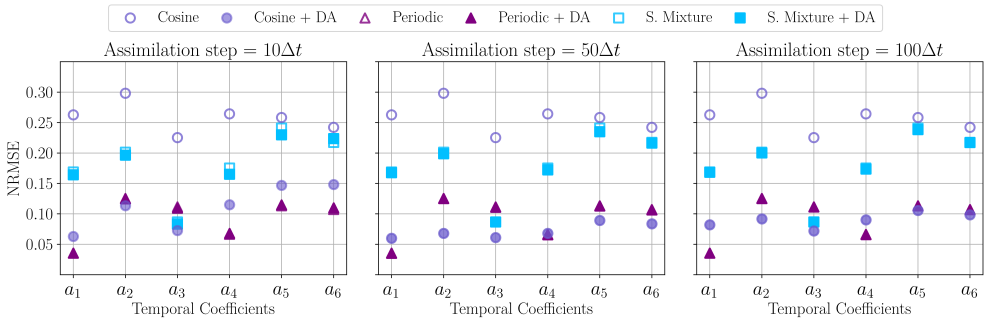


Fig. 9: NRMSE of the temporal-coefficient forecasts before and after data assimilation for the Cosine, Periodic, and Spectral Mixture GPR kernels using assimilation steps of $10\Delta t$, $50\Delta t$, and $100\Delta t$.

In general, reducing the assimilation interval from $50\Delta t$ to $10\Delta t$ does not lead to a systematic reduction in NRMSE. Indeed, for certain coefficients, the error even increases slightly. Conversely, increasing the interval to $100\Delta t$ results in only a moderate degradation of the forecast, most notably for the Cosine kernel. These results highlight a trade-off between correcting the accumulation of forecast errors and incorporating new latent-space observations. More frequent updates may provide limited additional benefits when successive observations contain highly correlated information,

particularly in the presence of measurement uncertainty. Conversely, longer assimilation intervals allow forecast errors to accumulate over more extended periods before correction.

Table 1: NRMSE of the reconstructed physical fields from GPR forecast.

Kernel	**Without DA**	**DA: 10Δt**	**DA: 50Δt**	**DA: 100Δt**
Cosine	0.073	0.038	**0.038**	0.040
Periodic	0.040	0.040	**0.040**	0.040
S.Mixture	0.053	0.050	**0.052**	0.053

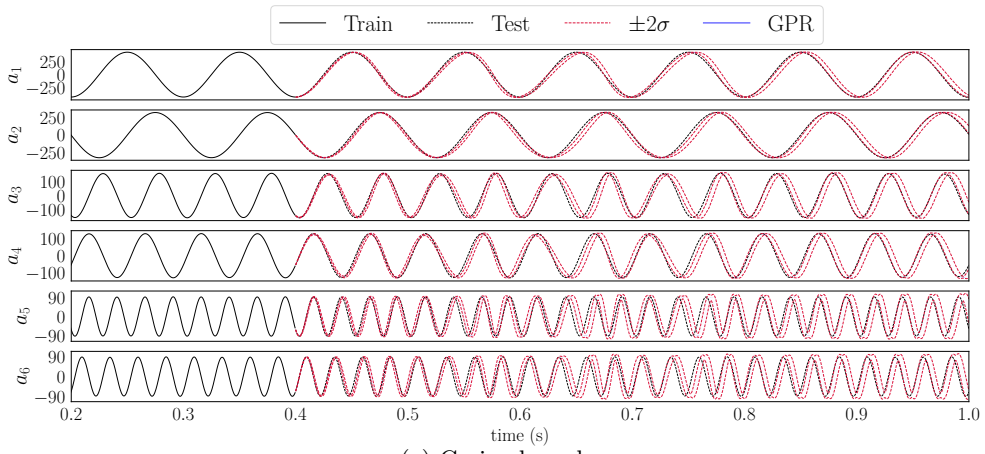
The reconstruction errors reported in Table 1 further confirm the limited sensitivity of the physical-field predictions to the assimilation interval for Periodic and Spectral Mixture GPR-based predictions. The Periodic kernel exhibits virtually unchanged reconstruction errors, remaining at approximately 0.04, while the Spectral Mixture kernel shows similarly minor variations, with NRMSE values between 0.050 and 0.053. For the Cosine kernel, EnKF assimilation reduces the reconstructed-field NRMSE from 0.073 to approximately 0.038, with values ranging from 0.038 to 0.040 across the different assimilation intervals. Hence, the Cosine kernel with assimilation performs as well as the Periodic kernel.

4.2 Numerical case

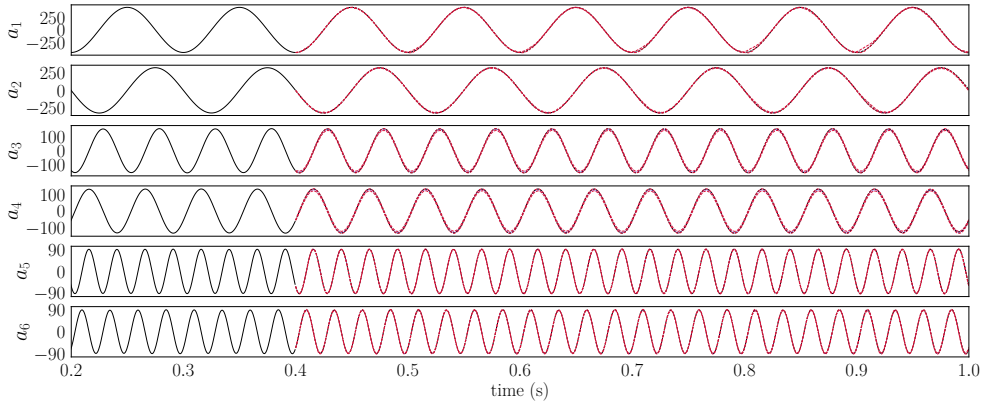
Following the methodology applied to the BIMER dataset, the laminar flame dataset was projected onto the mPOD basis, retaining six coefficients that describe the dominant temporal evolution of the system as they capture 94% of the total variance. The multi-frequency modulation imposed at the fuel inlet gives rise to multiple characteristic oscillatory scales, rather than a single dominant periodic response. The mPOD decomposition organizes these contributions into distinct latent trajectories, reducing the mixing of different temporal scales. These six coefficients therefore provided a compact, frequency-resolved representation of the flame dynamics and were subsequently used for the GPR forecasting and sparse-sensing analyses.

4.2.1 GPR perfomance

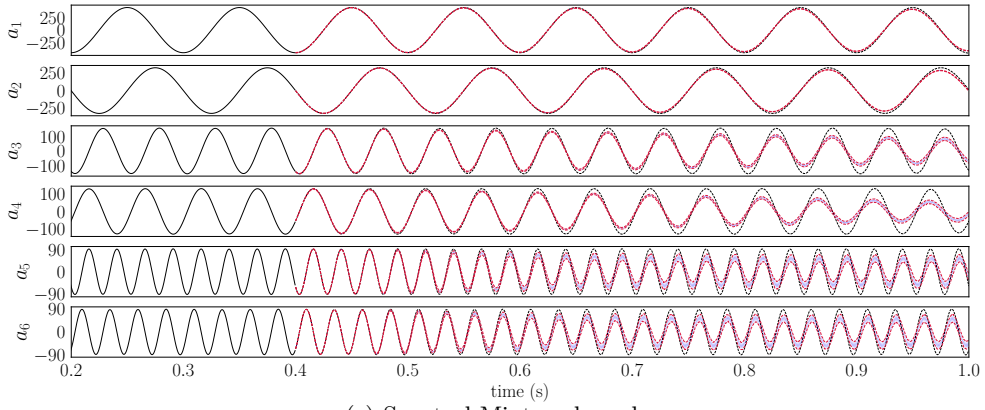
As for the BIMER case, the influence of the GPR kernel on the prediction of the mPOD temporal coefficients was assessed. As shown in Figure 10, the Periodic kernel accurately reproduces the temporal evolution of the six coefficients, preserving both the phase and amplitude throughout the forecast interval. In contrast, the Cosine kernel is primarily affected by an increasing phase mismatch, particularly for the higher-order coefficients, whereas the Spectral Mixture kernel generally captures the oscillatory timing more accurately but progressively underestimates the oscillation amplitude. These different forecasting behaviours are reported in Figure 11, where the Periodic kernel consistently yields the lowest NRMSE for all coefficients, followed by the Spectral Mixture kernel and finally by the Cosine kernel.



(a) Cosine kernel



(b) Periodic kernel



(c) Spectral Mixture kernel

Fig. 10: Forecast of the first 6 temporal coefficients using different GPR kernel formulations. The GPRs were trained from 0.1 to 0.4 s.

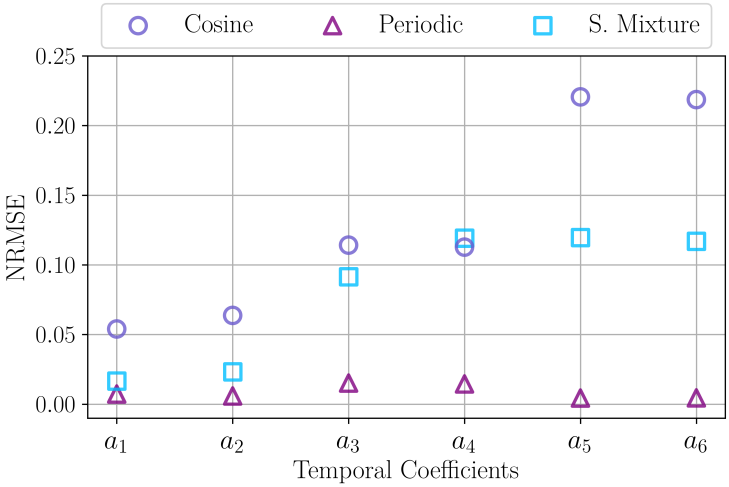


Fig. 11: NRMSE of the temporal-coefficient forecasts for different GPR kernels.

Table 2: NRMSE of the reconstructed physical fields from GPR forecast.

	Kernels		
Feature	**Cosine**	**Periodic**	**S. Mixture**
Temperature	0.023	**0.014**	0.017
CH_4	0.013	**0.010**	0.011
CO	0.034	**0.027**	0.030
CO_2	0.022	**0.014**	0.017
H_2	0.028	**0.021**	0.024
H_2O	0.020	**0.011**	0.014
N_2	0.012	**0.009**	0.010
O_2	0.020	**0.013**	0.016
OH	0.036	**0.030**	0.033

The differences observed in the temporal coefficients are also reflected in the reconstruction of the physical fields. Table 2 summarizes the NRMSE obtained for the reconstructed variables across the test dataset. Among the models, the GPR with Periodic kernel provides the lowest reconstruction error for all the variables, while the Spectral Mixture kernel yields intermediate errors that are generally close to those of the Periodic kernel. The Cosine kernel delivers the largest reconstruction errors for all variables. Interestingly, the differences among the kernels are less pronounced in the reconstructed physical fields than for some of the higher-order temporal coefficients. This can be explained by the different contribution of the mPOD modes to the reconstruction: the dominant low-order modes contain most of the retained energy and are

predicted relatively well by all three kernels, whereas the larger errors observed for the higher-order modes have a smaller impact on the reconstructed fields.

Figure 12 illustrates the reconstructions for temperature, CH_4 and OH at $t = 0.90$ s, well beyond the training interval, showing that the main flame structure and spatial features of the reacting fields are accurately captured by the Periodic GPR, indicating as well that the accuracy achieved in reduced space is effectively transferred to the high-dimensional representations.

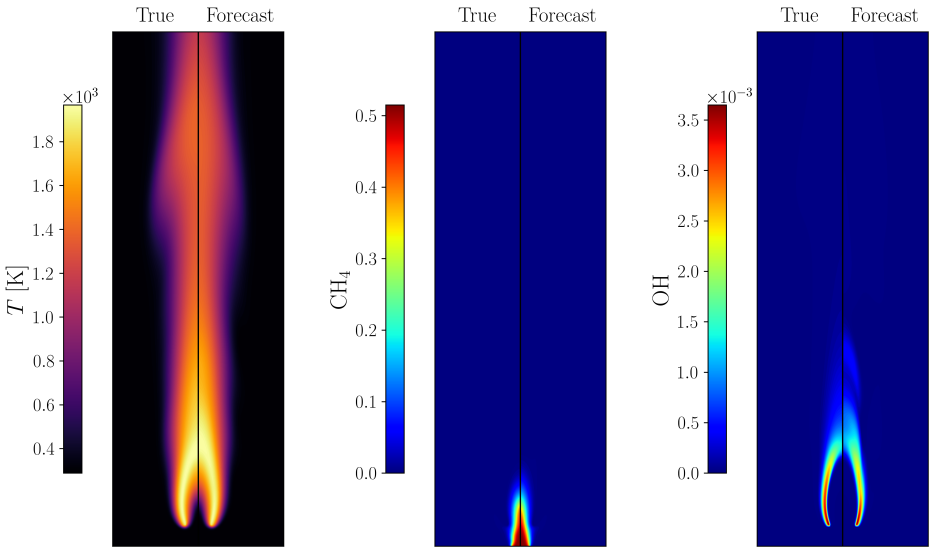


Fig. 12: Comparison between the reference and GPR-reconstructed fields at $t = 0.90$ s using the Periodic kernel GPR.

4.2.2 Sparse sensing and data assimilation

To evaluate the data assimilation framework under a more restrictive sensing scenario, only six values were selected at optimal locations provided by using QR pivoting. These locations were constrained to *numerical* measurements of temperature only through the Pysensors2.0 package [28], as in the BIMER case. In addition, a synthetic measurement noise was introduced by adding independent zero-mean Gaussian perturbations to the temperature observations. The noise standard deviation was set to 30% of the standard deviation of temperature values in the test dataset (where data assimilation was performed), corresponding to approximately 174 K.

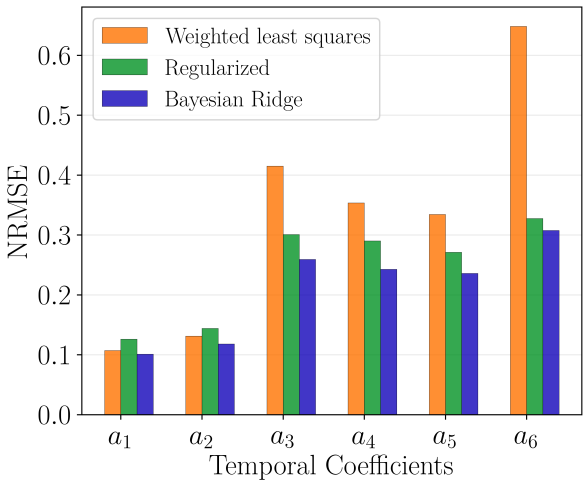


Fig. 13: Latent-state reconstruction error using six temperature values at optimal sensor locations and different inversion methods.

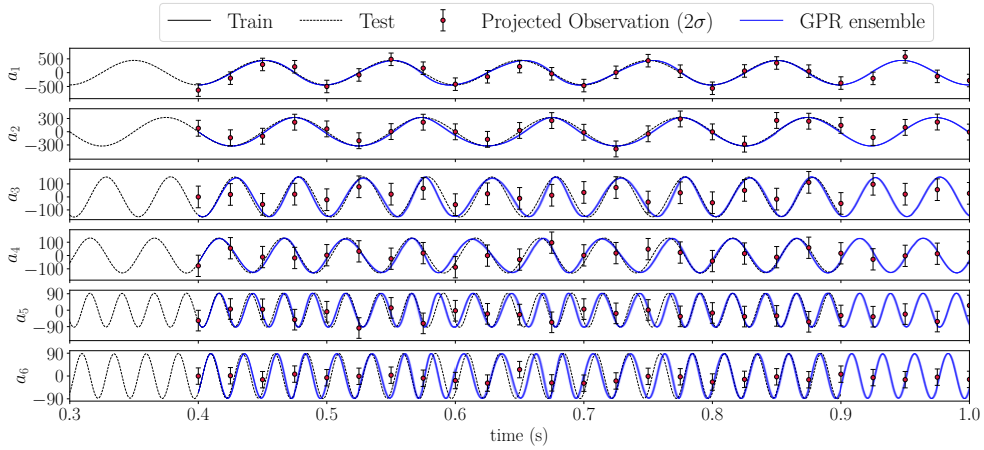


Fig. 14: EnKF correction of the Cosine kernel GPR forecast using sparse latent observations. The observations are assimilated every $50\Delta t = 0.025$ s.

As Figure 13 shows, the sparse and noisy temperature measurements provided sufficient information to estimate the corresponding latent state, despite the absence of samples from the other physical variables. The accuracy of the latent-state reconstruction depends on the inverse method used. Weighted least squares exhibits the largest errors particularly for the higher-error coefficients, whereas the regularized least squares and Bayesian ridge approaches provides substantially more accurate and

stable reconstructions. Among them, Bayesian ridge leads to the lowest overall error, indicating a greater robustness to measurement noise and limited sensing information.

Table 3: NRMSE of the reconstructed physical fields from GPR forecast.

		Cosine		
Feature	**Without DA**	**DA: 10Δt**	**DA: 50Δt**	**DA: 100Δt**
Temperature	0.023	0.065	**0.019**	0.026
CH_4	0.013	0.024	**0.012**	0.014
CO	0.034	0.082	**0.031**	0.038
CO_2	0.022	0.062	**0.019**	0.025
H_2	0.028	0.073	**0.025**	0.032
H_2O	0.020	0.061	**0.016**	0.023
N_2	0.012	0.023	**0.011**	0.013
O_2	0.020	0.061	**0.017**	0.023
OH	0.036	0.068	**0.034**	0.039

The latent observations reconstructed by Bayesian ridge were then assimilated into the Cosine kernel GPR forecast, which corresponds to the least accurate baseline model among the tested kernels. Figure 14 shows that the EnKF correction improves the predicted latent coefficients when assimilation is performed every $50\Delta t = 0.025$ s, reducing both the prediction uncertainty and the discrepancies in amplitude and phase over the forecast horizon.

The improvement in the predicted coefficients is also reflected in the reconstructed physical fields. Table 3 reports the NRMSE of the reconstructed physical fields from the Cosine kernel GPR forecast before and after the EnKF correction at different assimilation intervals. Among these values, $50\Delta t$ provides the best overall performance, consistently reducing the reconstruction error across all the variables, despite temperature being the only observed quantity. Unlike the experimental case, the reconstruction accuracy exhibits a clear dependence on the assimilation interval. This stronger sensitivity can be related to the more restrictive observation scenario considered here, where only six noisy temperature measurements are used to reconstruct the latent state, whereas the BIMER case relies on a spatially extended observation window. Consequently, the balance between forecast-error accumulation and observation uncertainty becomes more dependent on the assimilation frequency.

Increasing the assimilation interval to $100\Delta t$ results in errors higher than the baseline values for all the variables, indicating that the longer intervals between successive corrections allow prediction errors in the GPR forecast to accumulate. On the other hand, reducing the interval to $10\Delta t$ leads to even worse predictions, with the NRMSE values approximately twice those of the baseline case. This suggest that, under the relatively large measurement noise considered here, a more frequent assimilation repeatedly propagates uncertainty from the sparse latent observations into the forecast. Consequently, the correction may be influenced by measurements noise before the GPR forecast has accumulated sufficient error to benefit from a new update. These results indicate that the effectiveness of the correction depends on an appropriate

balance between limiting the accumulation of forecast error and avoiding excessively frequent updates from uncertain sparse observation.

Figure 15 further illustrates the effect of data assimilation in the reconstructed physical space. The corrected Cosine kernel GPR recovers the main features of the temperature, CH_4, and OH fields, achieving a reconstruction quality comparable to that obtained with the Periodic kernel, as shown in Figure 12.

Despite the relatively large synthetic temperature perturbations ($\sigma_{\text{noise}} \approx 174$ K), assimilating every $50\Delta t$ still improves the reconstruction of the complete thermo-chemical state, indicating that the proposed framework can remain effective under substantial random measurement noise. Although this noise level should not be directly equated with thermocouple measurement uncertainty, it provides a stringent test of the framework under degraded sensing conditions.

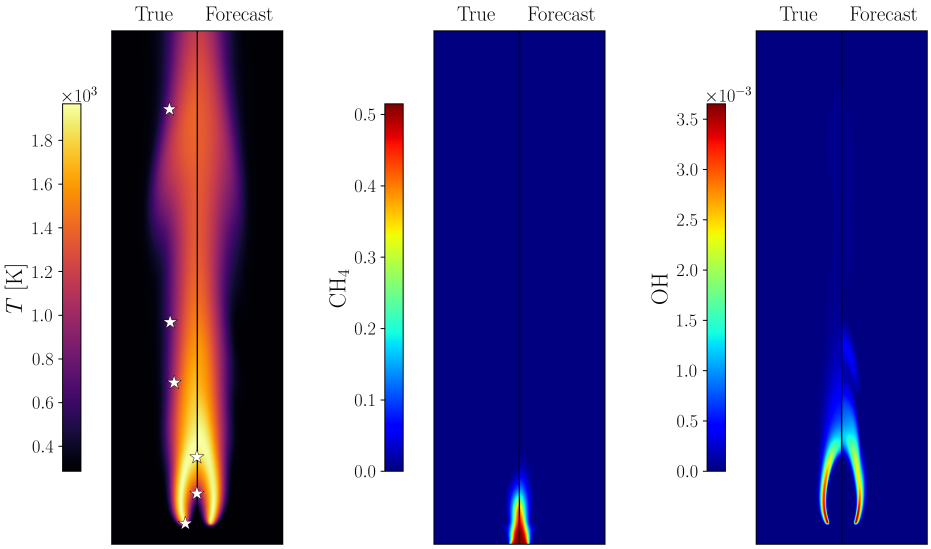


Fig. 15: Comparison between the reference and the reconstructed fields at $t = 0.90$s using the GPR with the Cosine kernel corrected through data assimilation every $50\Delta t$.

Although the assimilated Cosine GPR model does not fully outperform the Periodic one in terms of global reconstruction error, its performance becomes considerably closer. This result demonstrates that the proposed data assimilation strategy can partially compensate for deficiencies in the prior GPR forecast, even when only sparse and noisy temperature measurements are available. Overall, these results support that, even when the prior forecast is generated by a suboptimal kernel, sparse and optimally located observations of a single physical variable can be assimilated to improve the prediction, approaching the results of the best-performing forecast model.

5 Conclusions

This work presented an uncertainty-aware reduced-order modelling framework combining mPOD, Gaussian Process Regression (GPR), sparse sensing, and Ensemble Kalman filtering (EnKF) for forecasting combustion dynamics. The framework propagates uncertainty through the reduced-order prediction and assimilation processes, accounting for the confidence in both the probabilistic GPR forecast and the latent-state estimates obtained from sparse observations. Across the two test cases, the Periodic kernel in the GPR model provided the most accurate and robust forecast, highlighting the benefits of incorporating prior information about the characteristic oscillatory dynamics into the temporal covariance structure of the GPR model. For the pulsated laminar flame, the mPOD-based representation enabled accurate GPR forecasting of the latent dynamics despite the presence of multiple characteristic frequencies, by separating their contributions into distinct modes.

A systematic assessment of sparse-sensing configurations, inverse reconstruction methods, and assimilation frequency further clarified the respective roles of these components in the overall prediction performance. The BIMER results demonstrated that sparse-sensing performance depends not only on the number of observations but also on their spatial distribution, accessibility, and the conditioning of the inverse reconstruction problem. Restricting the sensing domain to a finite observation window increased the sensitivity to sensor placement and reconstruction method, with weighted least squares exhibiting the largest variations across configurations. Regularized least squares and Bayesian ridge regression provided more stable latent-state estimates under these constraints, while the selected 40×40 window offered a suitable compromise between spatial accessibility and reconstruction accuracy. Furthermore, increasing the number of available observations or the assimilation frequency did not systematically improve the forecast. These findings emphasize that the information content of the observations and their compatibility with the underlying forecast are more important than the quantity of observational data alone.

For the pulsed laminar flame, assimilating temperature values sampled at only six optimally selected sensor locations from the numerical solution, with substantial added noise, corrected the inaccurate prior forecast of the Cosine GPR model, and enabled reconstruction of the full thermo-chemical state with an accuracy approaching that obtained using the best-performing Periodic GPR model. Notably, the error reduction was comparable across the reconstructed thermo-chemical variables, even though only temperature was assimilated. This demonstrates how an appropriate reduced-order representation can exploit the coupling between state variables to propagate information from sparse temperature observations to the broader thermo-chemical state.

The present framework still presents some limitations. Data assimilation corrects the latent state but does not modify the underlying GPR dynamics trained offline. Consequently, dynamics insufficiently represented during training may still lead to forecast degradation between assimilation events. Future work will therefore focus on online adaptation of the reduced dynamical model, together with adaptive sensing strategies and validation under previously unseen forcing conditions. These developments will pave the way towards adaptive combustion digital twins that combine

efficient reduced-order prediction with data-driven correction and continuous learning from sparse observations.

Acknowledgements. The authors acknowledges the financial support from the European Union's Horizon Europe research and innovation programme and France 2030 funding, managed by the National Research Agency (ANR), as part of IA CLUSTER program, reference ANR-23-IACL-0003 - DATAIA CLUSTER.

Declarations

Funding

This work was supported by the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie COFUND (Grant agreement N°101127936 (DeMythif.AI)).

Competing Interests

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Ethics Approval and Consent to Participate

Not applicable.

Consent for Publication

All authors have given their consent for publication of this manuscript.

Materials Availability

Not applicable.

Code Availability

Available upon request.

Author contribution

Conceptualization: C.L., A.P., B.P., S.I.; Methodology: C.L., A.P., B.P., S.I.; Software: C.L., A.P., S.I.; Formal analysis: C.L., B.P., S.I.; Visualization: C.L.; Writing – original draft: C.L.; Writing – review and editing: A.P., B.P., S.I.; Supervision: B.P., S.I.; Funding acquisition: B.P, S.I.

References

[1] Hafeez, M.A., Procacci, A., Coussement, A., Parente, A.: Challenges and opportunities for the application of digital twins in hard-to-abate industries: a review.

Resources, Conservation and Recycling **209**, 107796 (2024) https://doi.org/10.1016/j.resconrec.2024.107796

[2] Pitsch, H.: Large-eddy simulation of turbulent combustion. Annual Review of Fluid Mechanics **38**, 453–482 (2006) https://doi.org/10.1146/annurev.fluid.38.050304.092133

[3] Chen, J.H.: Petascale direct numerical simulation of turbulent combustion—fundamental insights towards predictive models. Proceedings of the Combustion Institute **33**(1), 99–123 (2011) https://doi.org/10.1016/j.proci.2010.09.012

[4] Rowley, C.W., Dawson, S.T.M.: Model reduction for flow analysis and control. Annual Review of Fluid Mechanics **49**(Volume 49, 2017), 387–417 (2017)

[5] Taira, K., Brunton, S.L., Dawson, S.T.M., Rowley, C.W., Colonius, T., McKeon, B.J., Schmidt, O.T., Gordeyev, S., Theofilis, V., Ukeiley, L.S.: Modal analysis of fluid flows: An overview. AIAA Journal **55**(12), 4013–4041 (2017) https://doi.org/10.2514/1.J056060

[6] Taira, K., Hemati, M.S., Brunton, S.L., Sun, Y., Duraisamy, K., Bagheri, S., Dawson, S.T.M., Yeh, C.-A.: Modal analysis of fluid flows: Applications and outlook. AIAA Journal **58**(3), 998–1022 (2020) https://doi.org/10.2514/1.J058462

[7] Wan, Z.Y., Sapsis, T.P.: Reduced-space gaussian process regression for data-driven probabilistic forecast of chaotic dynamical systems. Physica D: Nonlinear Phenomena **345**, 40–55 (2017) https://doi.org/10.1016/j.physd.2016.12.005

[8] Ortali, G., Demo, N., Rozza, G.a.: A gaussian process regression approach within a data-driven pod framework for engineering problems in fluid dynamics. Mathematics in Engineering **4**(3), 1–16 (2021) https://doi.org/10.3934/mine.2022021

[9] Hafeez, M.A., Procacci, A., Coussement, A., Parente, A.: Constrained reduced-order modeling using bounded gaussian processes for physically consistent reacting flow predictions. Energy and AI **21**, 100554 (2025) https://doi.org/10.1016/j.egyai.2025.100554

[10] Aversano, G., Ferrarotti, M., Parente, A.: Digital twin of a combustion furnace operating in flameless conditions: reduced-order model development from cfd simulations. Proceedings of the Combustion Institute **38**(4), 5373–5381 (2021) https://doi.org/10.1016/j.proci.2020.06.045

[11] Mendez, M.A., Balabane, M., Buchlin, J.-M.: Multi-scale proper orthogonal decomposition of complex fluid flows. Journal of Fluid Mechanics **870**, 988–1036 (2019) https://doi.org/10.1017/jfm.2019.212

[12] Swischuk, R., Kramer, B., Huang, C., Willcox, K.: Learning physics-based reduced-order models for a single-injector combustion process. AIAA Journal **58**(6), 2658–2672 (2020) https://doi.org/10.2514/1.J058943

[13] Ni, C., Ding, S., Li, J., Chu, X., Ren, Z., Wang, X.: Projection-based reduced order modeling of multi-species mixing and combustion. Physics of Fluids **36**(7), 077168 (2024) https://doi.org/10.1063/5.0217845

[14] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, ??? (2006). https://doi.org/10.7551/mitpress/3206.001.0001

[15] Manohar, K., Brunton, B.W., Kutz, J.N., Brunton, S.L.: Data-driven sparse sensor placement for reconstruction: Demonstrating the benefits of exploiting known patterns. IEEE Control Systems Magazine **38**(3), 63–86 (2018) https://doi.org/10.1109/MCS.2018.2810460

[16] Lopez, C., Herrmann, B., Demarco, R., Escudero, F.: Sparse sensor placement and physics informed neural networks for temperature and velocity fields reconstruction in non-premixed laminar flames. Combustion Science and Technology **198**(11), 3763–3782 (2026) https://doi.org/10.1080/00102202.2026.2642891

[17] Rosati, V., Ariemma, G.B., Sorrentino, G., Ragucci, R., Joannon, M.: Sparse sensor placement for a cyclonic burner under mild combustion conditions. Flow, Turbulence and Combustion **116** (2026) https://doi.org/10.1007/s10494-026-00733-3

[18] Evensen, G.: In: Pinardi, N., Woods, J. (eds.) Sequential Data Assimilation for Nonlinear Dynamics: The Ensemble Kalman Filter, pp. 97–116. Springer, Berlin, Heidelberg (2002). https://doi.org/10.1007/978-3-662-22648-3_6

[19] Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. Journal of Geophysical Research: Oceans **99**(C5), 10143–10162 (1994) https://doi.org/10.1029/94JC00572 https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/94JC00572

[20] Labahn, J.W., Wu, H., Coriton, B., Frank, J.H., Ihme, M.: Data assimilation using high-speed measurements and les to examine local extinction events in turbulent flames. Proceedings of the Combustion Institute **37**(2), 2259–2266 (2019) https://doi.org/10.1016/j.proci.2018.06.043

[21] Donato, L., Kamal, M.M., Procacci, A., Cafiero, M., Sharma, S., Galletti, C., Coussement, A., Parente, A.: Integrating data assimilation and sparse sensing for updating a digital twin of a semi-industrial furnace. Proceedings of the Combustion Institute **40**(1), 105284 (2024) https://doi.org/10.1016/j.proci.2024.105284

[22] Donato, L., Galletti, C., Parente, A.: Self-updating digital twin of a hydrogen-powered furnace using data assimilation. Applied Thermal Engineering **236**, 121431 (2024) https://doi.org/10.1016/j.applthermaleng.2023.121431

[23] Procacci, A., Iavarone, S., Coussement, A., Parente, A.: Stochastic reduced-order modeling for the forecast of noisy dynamical systems. Proceedings of the Combustion Institute **41**, 105981 (2025) https://doi.org/10.1016/j.proci.2025.105981

[24] Wilson, A., Adams, R.: Gaussian process kernels for pattern discovery and extrapolation. In: Dasgupta, S., McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 28, pp. 1067–1075. PMLR, Atlanta, Georgia, USA (2013)

[25] Jolliffe, I.T.: Principal Component Analysis. Springer, ??? (2002)

[26] Lloyd, S.: Least squares quantization in pcm. IEEE Transactions on Information Theory **28**(2), 129–137 (1982) https://doi.org/10.1109/TIT.1982.1056489

[27] Klishin, A.A., Kutz, J.N., Manohar, K.: Data-Induced Interactions of Sparse Sensors Using Statistical Physics (2025). https://arxiv.org/abs/2307.11838

[28] Karnik, N., Bhangale, Y., Abdo, M.G., Klishin, A.A., Cogliati, J.J., Brunton, B.W., Kutz, J.N., Brunton, S.L., Manohar, K.: Pysensors 2.0: A python package for sparse sensor placement. Journal of Open Source Software **11**(118), 9265 (2026)

[29] Tipping, M.E.: Sparse bayesian learning and the relevance vector machine. Journal of Machine Learning Research **1**, 211–244 (2001)

[30] Evensen, G.: The ensemble kalman filter: Theoretical formulation and practical implementation. Ocean Dynamics **53**, 343–367 (2003) https://doi.org/10.1007/s10236-003-0036-9

[31] Renaud, A., Ducruix, S., Scouflaire, P., Zimmer, L.: Flame shape transition in a swirl stabilised liquid fueled burner. Proceedings of the Combustion Institute **35**(3), 3365–3372 (2015)

[32] Corrochano, A., D'Alessio, G., Parente, A., Clainche, S.L.: Higher order dynamic mode decomposition to model reacting flows. International Journal of Mechanical Sciences **249**, 108219 (2023) https://doi.org/10.1016/j.ijmecsci.2023.108219

[33] Smith, G.P., Golden, D.M., Frenklach, M., Moriarty, N.W., Eiteneer, B., Goldenberg, M., Bowman, C.T., Hanson, R.K., Song, S., Gardiner Jr, W., et al.: GRI-Mech 3.0, 1999 (2011)

[34] Corrochano, A., Freitas, R.S.M., López-Martín, M., Parente, A., Le Clainche,

S.: A predictive physics-aware hybrid reduced order model for reacting flows. International Journal for Numerical Methods in Engineering **126**(11), 70058 https://doi.org/10.1002/nme.70058