In Situ Distribution Guided Analysis and Visualization of Transonic Jet Engine Simulations

Soumya Dutta, Chun-Ming Chen, Gregory Heinlein, Han-Wei Shen, Member, IEEE, and Jen-Ping Chen

Abstract—Study of flow instability in turbine engine compressors is crucial to understand the inception and evolution of engine stall. Aerodynamics experts have been working on detecting the early signs of stall in order to devise novel stall suppression technologies. A state-of-the-art Navier-Stokes based, time-accurate computational fluid dynamics simulator, TURBO, has been developed in NASA to enhance the understanding of flow phenomena undergoing rotating stall. Despite the proven high modeling accuracy of TURBO, the excessive simulation data prohibits post-hoc analysis in both storage and I/O time. To address these issues and allow the expert to perform scalable stall analysis, we have designed an *in situ* distribution guided stall analysis technique. Our method summarizes statistics of important properties of the simulation data *in situ* using a probabilistic data modeling scheme. This data summarization enables statistical anomaly detection for flow instability in post analysis, which reveals the spatiotemporal trends of rotating stall for the expert to conceive new hypotheses. Furthermore, the verification of the hypotheses and exploratory visualization using the summarized data are realized using probabilistic visualization techniques such as uncertain isocontouring. Positive feedback from the domain scientist has indicated the efficacy of our system in exploratory stall analysis.

Index Terms—*In situ* analysis, rotating stall analysis, Gaussian mixture model, incremental distribution modeling, feature analysis, high performance computing, collaborative development

1 Introduction

Recent advancements of parallel computing capabilities have enabled aerodynamics scientists to study the phenomenon of rotating stall in great detail by performing high resolution numerical simulations. Rotating stall initiates from local airflow disturbances among the engine compressor blades, but grows quickly to become destructive to the engine. It is challenging to predict this event since the signs of stall inception are subtle and non-trivial to detect robustly. Numerous efforts have been made in the past decades to understand this phenomenon in detail. Recently, a computational fluid dynamics simulator TURBO [11, 12] has been developed in NASA, which is capable of accurately modeling the behavior of transonic compressors throughout their operational range, i.e. choke to stall. However, the computational cost and the amount of data produced from a single simulation is quite significant. Traditional post-processing analysis utilizing raw data cannot be readily applicable since storing all the raw data is not a viable option. This is because of the bottleneck stemming from I/O, compared to the ever increasing computing speed. Hence, exploration and visualization of such large scale data poses significant challenges.

The goal of this work is to analyze and visualize the spatiotemporal evolution of rotating stall in a jet engine simulation. More specifically, the expert wants to identify the blade passages and time step ranges when stall is imminent. Detection of early signs of stall is critical for the analysis since it enables the expert to perform exploration with a focus on the relevant data. In addition, the expert wants to concentrate on the transition of the simulation from a stable condition to an unsteady state around the identified time step ranges which leads to engine stall. An important point to note is that, for some operational conditions, it is unclear whether the simulation is going to stall. Therefore, it is often necessary to run the simulation for several days in supercomputers which will produce a data set that is too large to store. Thus, the expert wants to have a significant storage reduction in their data so that scalable post analysis is possible. In order to study rotating stall inception, the expert also wants to take a multifaceted approach requiring several variables. Since stall is generally considered as an abnormal behavior in airflow through compressor blades, the expert is interested to look at the stall phenomenon by exploring potential anomalous regions in the data over space and time. Finally, visualization techniques are required which can be used to validate the hypotheses and formulate new reasoning for a better understanding of rotating stall.

In situ analysis, i.e., in-place analysis of data while it still resides in memory, presents an attractive option to remedy many of the aforementioned issues. It helps to avoid slow data output to secondary storage by performing analysis while data are produced. Furthermore, such in-place analysis enables us to apply a suitable data triage during the simulation time for extracting and preserving important information compactly via data summarization techniques [1, 13, 44]. However, it is to be noted that the amount of work we can do *in situ* is also constrained in terms of time and storage space since overburdening the simulation is undesired. Some stall analysis tasks requiring data from the global spatial domain are thus not preferable to run *in situ*, since it will impose additional data communication and memory consumption among processes in the distributed memory environment. Therefore, choices between tasks that can run *in situ* or should be deferred to the post analysis are important to make.

In this work, we propose a scalable method for rotating stall analysis which exploits the advantages of *in situ* analysis and facilitates exploratory post analysis. It allows us to tame the challenges posed by the extreme scale data and perform exploration in an efficient manner. Since rotating stall is in general characterized by local disturbances in airflow, it is non-trivial to have a precise descriptor for their detection. Use of statistical methods in such scenario has shown promising results [9, 19]. Analysis using probability distributions have benefited many visualization tasks along this line in the recent past [15, 21, 24, 29, 46]. In this work, we use probability distributions to model local statistical properties of the data and analyze the spatiotemporal distribution variations to identify stall impacted regions. To efficiently and compactly capture the statistical data properties, we employ an incremental learning scheme to model time-varying data distributions in the form of Gaussian mixture models (GMM) [15, 40, 45] during the simulation. We show that storing data in this summarized

- *Soumya Dutta, Chun-Ming Chen, and Han-Wei Shen are with the GRAVITY research group, The Department of Computer Science and Engineering, The Ohio State University. E-mail: dutta.33, chen.1701, shen.94@osu.edu.*
- *Gregory Heinlein and Jen-Ping Chen are with The Department of Mechanical and Aerospace Engineering, The Ohio State University. E-mail: heinlein.29, chen.1210@osu.edu.*

Manuscript received 31 Mar. 2016; accepted 1 Aug. 2016. Date of publication 15 Aug. 2016; date of current version 23 Oct. 2016.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2016.2598604

representation enables flexible post analyses based study of rotating stall. This is done through leveraging the existing and new visualization algorithms with the much needed capability of uncertainty quantification [7, 27, 30, 33, 34, 35]. For validating the suspected locations in spatial domain, we utilize uncertain isocontour algorithms. Positive feedback from the expert confirms the efficacy and benefits of the proposed method and also demonstrates the capability of *in situ* processing in analyzing extreme scale data sets in an effective way. Therefore, Our contributions in this work are fourfold:

1. We introduce a scalable *in situ* stochastic data modeling technique by taking advantage of an incremental learning scheme of GMMs, which helps us to extract the important statistical properties of the data in simulation time in a compact format.
2. We present a novel distribution guided rotating stall exploration technique which exploits both spatial and temporal nature of stall by analyzing the statistical information of data stored during *in situ* processing.
3. We use a comparative visualization technique to conduct anomaly pattern analysis using multi-variables to obtain new insights about rotating stall.
4. Finally, we make use of uncertain isocontouring and other spatial visualization techniques to allow the expert to explore the stall analysis results in spatial domain for validating the hypotheses.

The rest of the paper is organized as follows: In Section 2 we present a discussion of the related works. Background and motivation of this work is discussed in Section 3 and the requirements are listed in Section 4. Section 5 presents the *in situ* data modeling scheme and in Section 6 we discuss the spatiotemporal anomaly based stall analysis in detail. Section 7 depicts the visualization techniques used in this work. In Section 8 the details of the *in situ* implementation are depicted. We demonstrate the results obtained by our method in Section 9 followed by performance study in Section 10. Finally, we conclude our work in Section 11 by highlighting several possible future directions.

2 Related Works

***In situ* processing, analysis, and visualization.** The necessity of *in situ* analysis is becoming more prominent as the size of data output from high-resolution simulations is out-pacing post-processing and visualization capabilities. One of the early attempts of *in situ* visualization was made by Haimes [17] to visualize large unsteady data sets. For enabling *in situ* capability in Paraview, Fabian et al. proposed CATALYST library [16]. Similarly, run-time visualization with LibSim using VisIt was introduced by Whitelock et al. [48] and in another work Lofstead et al. added ADIOS as an *in situ* visualization framework [28]. Vishwanath et al. enhanced simulation time data analysis by proposing GLEAN [42]. A zero copy data structure was introduced by Woodring et al. [50]. *In situ* eddy analysis in ocean simulation models was demonstrated by Woodring et al. [51]. Yu et al. enabled high quality *in situ* visualization of combustion data in their work [52].

However, visualization tasks which require exploratory data analysis, verification, and validation by adding experts in the loop can not be always done using traditional *in situ* approaches. Flexible post analysis will still be needed for scientific discovery. Hence, a new *in situ* based paradigm is emerging where the task specific important simulation data is massively reduced via *in situ* processing and flexible scalable post analysis is done on the reduced data [13]. The visualization community has begun to embrace this new paradigm and have proposed several such schemes [25, 44]. A sampling based method for interactive visualization of cosmology data was used by Woodring et al. [49]. Ahrens et al. adopted an *in situ* image based approach [1] for feature exploration during post analysis. In this work, we extend the methods of *in situ* analysis through statistical summaries of the data by means of scalable and compact probabilistic modeling to reduce data size and allow for user driven exploration.

Distribution-based data summarization and analysis. In query-driven visualization, Lundstrom et al. [29] used local histograms for transfer function design. Johnson and Huang [21] proposed a feature extraction method based on fuzzy matching of the user-queried frequency distribution and local distributions. Wei et al. [46] presented efficient local histogram search using bitmap indexing. Efficient range distribution query algorithms using integral histograms [9] and wavelet transforms [24] yield valuable statistical information on mean, variance, information entropy etc. Recently, GMM-based feature extraction has received increasing attention due to its compact representation and the close relation to clustering. Wang et al. [45] used multidimensional GMMs for transfer function design in time-varying datasets. Dutta and Shen [15] proposed fuzzy feature tracking based on block-wise GMMs. For large data summarization, Thompson et al. [41] presented the idea of hixel, which stored a histogram per data block to preserve uncertainty information due to data down-sampling. The authors showed the advantages of storing hixels including approximating the topological structures and extracting fuzzy isosurfaces, which provide more informative results. Liu et al. [27] presented similar idea but compacted the distribution representation by GMMs for stochastic volume rendering on the GPU. Our *in-situ* data summarization uses similar idea by Liu et al. to stores block-wise distributions in GMMs and generates spatial distribution datasets, where each spatial location presents a collection of values for the same variable [30].

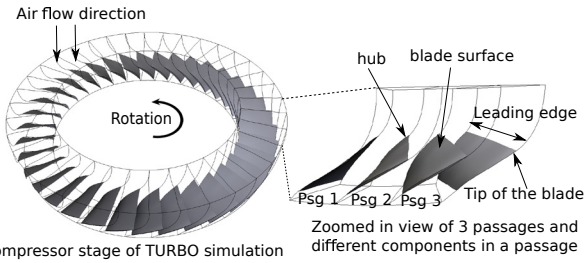


Figure 1: A diagram of the compressor stage of TURBO simulation and a zoomed in view of a blade passage.

Uncertainty visualization of spatial distribution datasets. Brodie et al. [7] and Bonneau et al. [5] recently provided thorough reviews. In visualizing spatial distribution datasets, Kao et al. [22], Luo et al. [30] and Potter et al. [35] visualized 2D distribution datasets by displaying statistical summaries such as means, standard deviations and skews in color, height field, or glyphs. Potter et al. [34] proposed summary plots which extend box plots with moments and histograms in higher dimension visualizations. To visualize 3D distribution datasets, flickering the color according to the distribution samples is used in volume rendering [27, 41]. Uncertain isosurface extraction provides further understanding of data at a specific isovalue. Pöthkow et al. [32] computed the level crossing probability of adjacent points, which was extended to computing cell-wise level crossing probability [33]. Athawale et al. [2] devised closed-form computation of level-crossing probabilities for nonparametric distribution datasets.

3 Application Background and Motivation

In this section, we describe the necessary background and motivation of this work. We further concentrate on the requirements set by the expert and provide an overview of our approach for solving them.

3.1 Background Concepts of Rotating Stall

Detection of rotating stall in compressor stages has been an important problem for aerospace scientists for a long time. Failure to detect stall has dire consequences in engine's functionality and can lead to permanent engine damage. Therefore, the scientists have always sought after techniques that can detect the sign of rotating stall as early as possible so that corrective control methods can be applied. Rotating stall initiates as intermittent local flow separation on the turbine blade surface, often caused primarily by fluid instabilities around the tip region of a blade. These regions initially contain small *bubble-like* blockages which hinder normal airflow through the passages. If the blockages increase over time and become persistent, they are characterized as *stall*

cells. By detecting and studying the local flow abnormalities in the early stages, further understanding of rotating stall formation can be gleaned. To study the rotating stall phenomenon in detail and understand how it develops over time, scientists from NASA have developed a high resolution CFD simulator TURBO [11]. It has been validated by previous works [11, 18] that TURBO is able to model the rotating stall with high resolution and hence provide detailed knowledge about the phenomenon of rotating stall. The rotor in this configuration consists of 36 blade passages, as shown in left sub-image of Figure 1. In order to view the structure of the passages a zoomed in view of a subset of the full rotor is depicted on the right sub-image of Figure 1. In this figure we highlight the different domain specific components of a blade passage.

3.2 Limitations of Current Stall Analysis Approaches and Motivation of Our Work

The motivation of our work stems from the limitations of existing stall analysis approaches which are twofold in nature: **(1)** The limitations that arise from excessive storage requirements, I/O bottleneck, and prolonged post-processing time; **(2)** The limitations of existing stall analysis techniques. Next we discuss each of these points briefly and justify the necessity of our work.

A full annulus simulation of TURBO consisting of 4 revolutions of the compressor stage takes around a day to finish in a supercomputer and produces around 20 TBs of raw data. Processing, handling, and analyzing such scale of data is posing a significant obstacle to the domain scientists. Therefore, a growing need of scalable and efficient analysis methods has become prominent. The bottleneck coming from I/O, and the cost of moving a huge amount of data from supercomputers to the local processing machines becomes prohibitive as the data size grows. Furthermore, traditional post-hoc stall analysis methods require longer time for such large data to produce any useful results.

Among the most utilized existing stall analysis techniques are mass flow rate and pressure probe observations. **Mass flow rate** is a measure of the air mass that flows through the compressor stage per unit time [18] and is defined as: $\dot{m} = \rho \vec{v} \cdot \vec{A}$, where ρ is the density, $\vec{A}$ is the area vector of a flow path cross section of the inlet or exit of the compressor, and $\vec{v}$ is the flow velocity. Mass flow rate in stable condition remains constant over time, however, when rotating stall happens, it drops rapidly. Unfortunately, the event is only observable as rotating stall is occurring and thereby cannot be used as a precursor for stall.

Analysis using **pressure probe** readings [14, 31] to capture pressure variations at fixed locations over time has a better potential of detecting earlier signs of stall. This technique employs pressure probes at the engine casing circumferentially. When a rotating instability passes through the probed locations, the pressure reading of the probes can fluctuate and by observing the time-varying patterns of such pressure probes, rotating stall can be detected. However, it is non-trivial to find appropriate probing locations and usually the pressure probe based methods only use a few probes to detect the stall phenomenon which does not provide detailed spatial information.

A recent work by Chen et al. [10] has shown the potential of the statistical anomaly based analysis in the detection of rotating stall. By performing the pressure anomaly analysis, it was shown that the subtle signs of rotating stall can be detected in much earlier time steps. However, as mentioned above, such a post-processing based analysis requires a long preprocessing time including data transfer and loading time. Furthermore, to reduce the storage cost, time steps were skipped in the simulation output. Therefore, even though their method has demonstrated a great potential in stall analysis, such post-hoc workflow will not scale well as the data size grows.

4 Domain Specific Requirements and Overview of Our Approach

The aforementioned problems of current stall analysis approaches have led us to design a new approach with a list of domain specific requirements. Below we list those requirements first and then present an overview of our approach for solving them.

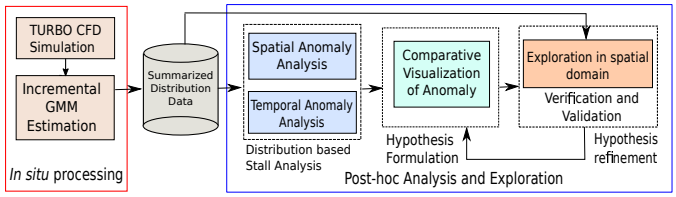


Figure 2: A schematic diagram of the proposed analysis method.

4.1 Requirements

Following are the requirements that have been identified:

1. Since the expert wants to have significant reduction in data size and yet preserve the important information which can be analyzed flexibly, an *in situ* analysis seems suitable.
2. Since the stall phenomenon is characterized by locally unstable regions, a spatial region based analysis will be more suitable. This will reduce the processing cost and yet detect the desired regions. Furthermore, apart from the spatial anomaly, the expert is also interested in finding anomalous airflow behavior in temporal domain and hopes to verify both analysis methods in capturing the early signs of stall.
3. A majority of the previous works have focused on observing the variation of pressure for stall analysis. The expert desires to look at other variables as well. It is hypothesized that the entropy values are also closely related to the stall cells, so the domain expert wants to compare the effectiveness of pressure and entropy as stall indicator variables.
4. The reduced data output should be used to render the results in spatial domain for exploration, verification, and validation.

4.2 Overview of Our Approach

To fulfill the above requirements set by the expert, we provide a new rotating stall analysis approach which integrates *in situ* data summarization with distribution-based analysis and visualization techniques. Figure 2 shows a schematic diagram of our proposed method. To tame the extreme size of the simulation output, during the simulation we take advantage of *in situ* processing to summarize the important data into GMM distributions. The *in situ* processing outputs spatial distribution datasets in a much smaller data size than the raw data, which enables efficient post-hoc analysis and visual exploration. To identify the stall suspected regions, we employ spatial and temporal anomaly detection methods on the stored distribution data, which measure statistical variations among GMMs over space and time. By studying the spatial and temporal anomaly results of multi-variables through comparative visualization, the expert can analyze the evolution of rotating stall and identify the locations where stall is initiated. Finally, to investigate the detected phenomena in spatial domain for validation and verification of the hypotheses, we allow the scientist to visualize the spatial distribution data with uncertain isosurfaces. In the following, we first present the *in situ* incremental GMM estimation algorithm and then discuss the distribution guided anomaly analysis in detail.

5 *In Situ* Incremental Probabilistic Data Modeling

In this section we provide a detailed description of the proposed *in situ* probabilistic data modeling technique. The features of stall to be identified are characterized by local disturbances in airflow, known as *stall cells*. Due to the complexity of the stall phenomena, there is no precise descriptor available for a stall cell in the literature. Since our goal is to summarize the data and identify such regions which show statistically anomalous behavior, a simple down sampling scheme will not work well because, such a scheme will fail to preserve the necessary local statistical properties of the data. Instead, a local distribution based data modeling in this scenario presents a suitable option [15, 29]. We model the data using probability distribution functions and adopt a block-wise approach for representing the local statistical signature of data for each data block. With such a data model, we can efficiently

estimate the data variation in local regions by comparing the distributions over space and time and quantify the possibility of a region (i.e. a data block) to be anomalous. Note that a block-wise approach meets the requirement of the local region based exploration mentioned earlier and also has been advocated suitable for analyzing large scale time-varying scientific data sets [15, 43].

5.1 Mixture of Gaussians for Compact Distribution Representation

The distribution model in our case needs to be compact and storage efficient such that post-processing using such data products scales well and also allows flexible exploration. Popular non-parametric distribution estimator Kernel Density Estimation (KDE) requires high storage and computational costs. Another estimator, histogram, can be computed quickly but they also suffer from higher storage requirement. In the parametric domain, using a Gaussian distribution to estimate the probability density of a data block can reduce the storage significantly but the assumption of Gaussianity may not be accurate when the underlying data is multi-modal. Therefore, we adopt mixtures of Gaussians (GMM), a parametric distribution modeling scheme which is storage efficient and can be computed incrementally over time. Since several Gaussians are combined to estimate the underlying data distribution, no prior assumptions about the distribution are made. Furthermore, based on Gaussian properties, GMMs also allow efficient computation of numerous statistical measures of the data which makes it an attractive choice for our use. Formally, the probability density function $p(x)$ of a GMM for a random variable X is expressed as:

$$p(x) = \sum_{j=1}^{K} \omega_j * \mathcal{N}(x|\mu_j, \sigma_j) \tag{1}$$

where K is the number of Gaussian components. ω_j, μ_j and, σ_j are the weight, mean, and standard deviation for the j^{th} Gaussian component respectively. The sum of weights in the mixture, $\sum_{j=1}^{K} \omega_j$, is always equal to 1. Next, we present the details of an incremental Gaussian mixture model learning algorithm for *in situ* distribution estimation.

5.2 Scalable Incremental Estimation of Gaussian Mixture Model

Here we describe the *in situ* estimation of time-varying GMMs using an incremental mixture model learning algorithm. Traditional Gaussian mixture model learning algorithms use Expectation Maximization (EM) to maximize the likelihood function [4] for estimating the parameters of the mixture model. However, such a technique follows an iterative approach and requires longer time to complete. Since our target is to estimate the GMM *in situ* in a scalable way without impacting the overall simulation time too much, we advocate for an alternative incremental learning scheme for mixture of Gaussians proposed in [40]. Such an incremental updating based algorithm for estimating GMMs was shown to be computationally fast and suitable for modeling large scale time-varying data sets [15, 45].

We model the statistical behavior of each local data block using mixtures of Gaussians in this work. Such a local region based probabilistic model runs *in situ* and as the data for new time step is produced, the incremental model for each block updates the distribution parameters using the new data to represent the current distribution for the block. At each time step, for each block, every new data point is checked against the existing K Gaussian distributions in the model. A match is identified if a data point lies within the 2.5 standard deviation of a Gaussian in the model. When multiple matches are found, then the best matched Gaussian is picked. If none of the K Gaussians match the current data point, then the least probable distribution in the model is replaced with a new Gaussian with the current data value as its mean, an initial high standard deviation, and a low weight [40]. The weights at time t for the j^{th} mixture are adjusted as:

$$\omega_{j,t} = (1-\gamma)\omega_{j,t-1} + \gamma \mathscr{I}_{j,t}, \quad j \in \{1, 2...K\} \tag{2}$$

γ is called the learning rate and the value of $\mathscr{I}_{j,t}$ is 1 for the distribution with the best match and 0 otherwise. After the adjustment, all the weights are normalized again for maintaining consistency. The μ and σ parameters for the unmatched distributions remain the same, however for the matched distribution they are updated as:

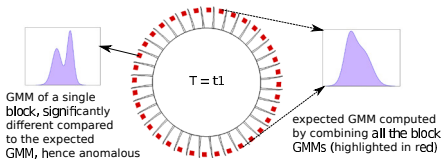


Figure 3: Illustration of spatial anomaly detection method using GMM distributions over space.

$$\mu_{j,t} = (1-\gamma)\mu_{j,t-1} + \gamma x_{j,t} \tag{3}$$

$$\sigma^2_{j,t} = (1-\gamma)\sigma^2_{j,t-1} + \gamma(\mu_{j,t} - x_{j,t})^2 \tag{4}$$

Once we have observed all the points for a block in the current time step, the GMM will give us the updated distribution. It is evident that the model adapts to the new data since it adds or removes Gaussians from the existing model as required. Since the incremental GMM estimation algorithm requires an initial parameter set to start with, we employ the traditional expectation maximization based learning algorithm for the first time step only to get an initial estimation of the GMM parameters for each block and then the incremental update algorithm is invoked for modeling GMMs for all the future time steps.

6 Post-hoc Probabilistic Stall Analysis using Mixture of Gaussians

In this section, we present the stall analysis techniques which exploit the statistical information summarized in the form of GMMs that were generated *in situ*. Since the domain expert describes stall cells as local instability in the airflow, in this work, we have focused on identifying such instability by characterizing them as local statistical anomaly in the data. In a stable condition, all the blade passages of the compressor are expected to be *axisymmetric*. Hence, a specific local region (i.e. a block) relative to all the blade passages would behave symmetrically. Furthermore, the observed values of simulation variables such as pressure, and entropy at a local region in a specific blade passage would be similar in future time steps as well. Based on these two key observations, a classification for anomalous regions over both space and time can be achieved; where quantification of a region being anomalous is defined by the amount of statistical dissimilarity in the same region of other passages over space and time. In the following, we first describe our spatial anomaly based analysis and then introduce a temporal anomaly measure for rotating stall analysis.

6.1 Spatial Anomaly Guided Stall Analysis

In our *in situ* distribution guided analysis, we aim at detecting local anomalous regions i.e. the data blocks which are expected to contain stall cells. In a recent work, a point based spatial anomaly detection method was proposed by Chen et al. [10] and was shown to have high potential for identifying instability which indicated early formation of stall cells. The method used the pressure variable for the analysis and exploited the property of axisymmetry among blade passages.

In our work we advocate for a local region based analysis for the detection of stall cells rather than a point based analysis because the point based method requires a whole domain analysis on the raw data, which is too expensive to perform post-hoc or compute *in situ*. Furthermore, the expert believes that, since the stall cells form a spatial region, a local region based analysis is more suitable. Following these thoughts, in this work, block-wise local distributions are modeled and stored in the form of GMMs for efficient post analysis. To detect spatial anomaly, we first group the GMMs coming from the same relative location in each passage. As illustrated in Figure 3, each group contains 36 blocks (highlighted in red), i.e. 36 GMMs, coming from 36 blade passages in the rotor. Thus, our goal of identifying the spatial anomaly among axisymmetric regions is essentially to find outliers among a group of GMM distributions.

To identify a GMM that consists of abnormal values among a group of GMMs, our approach first estimates the *expected* distribution as a basis for all the GMMs in the group to compare with. If any of the 36 GMMs is sufficiently different from the *expected* distribution, it is regarded as an outlier and the corresponding block in the physical space is reported anomalous. We define the *expected* distribution as the average of the probability density functions of the GMMs in the group. This is equivalent to computing the combined distribution from all samples in the group of GMMs. In this way the major value distribution is attained and the effect of outlier values to the *expected* distribution, if present, is reduced. The *expected* GMM is a new GMM distribution consisting of the Gaussians from all the input GMMs with normalized weights.

After the *expected* GMM is formed for a group, we compare it with each GMM in the group using the Earth Mover's Distance (EMD). The EMD is a distance measure defined by the minimal ground transport effort to match two distribution shapes. EMD has been widely used in pattern matching and image analysis [23, 38], as well as to compare probability distributions in uncertain data [37]. Besides its robustness against noise [26], EMD's measuring of ground transportation is able to capture an outlier if its value deviates from the majority of the distribution, which is particularly desired in our anomaly detection study. To compute the EMD for 1D distributions, we use the *match distance* as the ground distance [36] since the EMD can thus be efficiently computed by the absolute difference between the cumulative distribution functions (CDF) of the distributions [47]:

$$EMD(X,Y) = \int_{-\infty}^{\infty} |F_X(x) - F_Y(x)|\,dx \tag{5}$$

Here $F_X(x)$ is the CDF of the distribution X at x. The CDF of a GMM can be simply computed by the weighted sum of the CDFs of the individual Gaussians. We numerically approximate the integral of CDF difference by the trapezoidal rule. After the EMD is computed, a user specified fixed threshold is used to extract the blocks with outlier GMMs. We apply the above method for all groups of blocks per time step and mark the blocks identified as spatial anomaly.

6.2 Temporal Anomaly Guided Stall Analysis

GMM based spatial anomaly measure presented above can have a potential limitation. During a stall developing phase the anomalous regions might propagate to the majority of the blade passages and in that case, the outlier values will dominate the *expected* GMM and the spatial anomaly test may not be able to detect them as anomaly anymore. Therefore, in this section, we extend the anomaly based analysis to temporal domain to remedy such a situation. Since the GMMs for each local region (i.e. the data block) gets updated as data from new time step is observed, it is hypothesized that in a stable state, the GMMs for a block will not change significantly over time. Therefore, by observing the temporal dissimilarity between the GMMs for the same block over time, temporal anomaly can be computed for each block. Since temporal anomaly is computed by looking at each block individually over time, even if majority of the blocks get affected by the stall, temporal anomaly method will still detect them as anomalous.

Another motivation to investigate the rotating stall using temporal anomaly comes from a domain specific hypothesis. The expert thinks that when the stall has fully developed, the temporal variation of data values inside the fully grown stall cells may become less apparent compared to that of the stall developing phase. Since temporal anomaly will measure the instability of data values inside a block by comparing the block GMMs over time, it is expected that the degree of temporal anomaly for a block contained in a developed stall cell will be small. Such a behavior of stall cells can be investigated by observing the pattern of detected temporal anomaly in an appropriate time window.

In Figure 4, we present the idea of temporal anomaly using an illustrative diagram. On the left side at time t1, we show that a specific block (highlighted in red) is selected for analysis and its GMM is estimated. Next the GMM of the same block coming from the same blade passage in the next time step is observed. If the similarity between

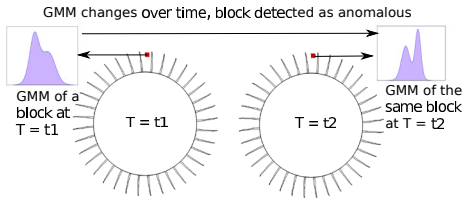


Figure 4: Illustration of temporal anomaly detection method using GMM distributions over time.

the two GMMs over time for this block is less than a preferred degree, then the block is classified as anomalous. As shown in Figure 4, the selected block at time t2 has a GMM which is quite different from its previous state and therefore the block is detected to be anomalous at time t2. To measure the similarity between the GMMs of a block over consecutive time steps, we have used the Earth Mover's Distance (EMD) which is introduced in Section 6.1. Therefore, applying this similarity based measurement to all the blocks over time, we estimate the chance of a block containing a temporal anomaly.

7 Visualization Techniques for Exploration and Verification of Stall Analysis Results

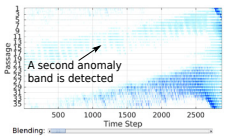
After the spatial and temporal anomalies are identified as flow instability, the next goal is to help the domain expert to relate the anomalies to the evolution of stall. Based on the requirements of the expert, we present a comparative chart as shown in Figure 5c that effectively shows spatiotemporal evolution of anomalies detected by different methods. Interesting time steps and regions are then selected and visualized in the physical space for verification of rotating stall and further exploration. It is to be noted that, in post analysis we do not have access to the raw data anymore, instead local distributions in the form of GMMs are available. Therefore, spatial visualization techniques that analyze probability distributions are employed in our work. Below we first describe the comparative anomaly chart as an overview visualization, followed by the spatial visualization techniques used for validation and exploration.

7.1 Comparative Visualization for Anomaly Pattern Study

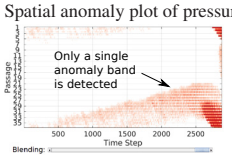
As the anomaly based stall analysis methods described in Section 6 estimate the chance of instability for all regions over time, it is important to provide such information to the expert through an overview showing the pattern and evolution of the detected anomalous regions over time. By investigating the trends of detected regions from the global view, the expert can quickly identify the blade passages and time step ranges for further examination in data domain. Moreover, a comparative visualization technique is applied for the expert to compare with anomalies detected in different variables for hypothesis verification.

The anomaly chart. According to the domain expert, stall cells that cause engine stall generally have the following properties: (1) They can exist and propagate across passages for a long time, and (2) They can grow in size to hinder normal airflow through the compressor. In the overview visualization, the expert is interested in how the detected flow instability propagates among passages instead of how it moves inside a passage. Therefore, a 2D heat map is used to visualize the anomaly detection results, where the Y axis on the chart represents the passage number and the X axis represents the time step, as shown in Figure 5. It is to be noted that, since in the physical domain the blocks in the compressor are organized in circular fashion, the anomaly chart also can be a polar plot which will correspond the actual structure, however, the expert has found that the 2D plot is more effective in showing the evolution of anomalous regions clearly. Additionally, the slant patterns in the 2D plot assist the expert to derive the speed of the rotating anomalous regions which would be difficult to measure in a polar plot.

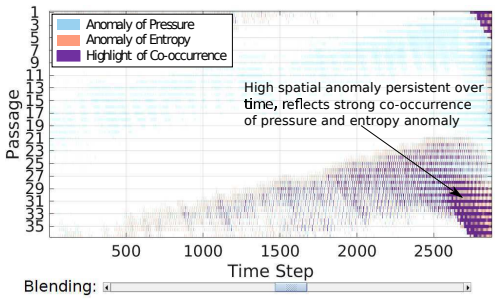
Each point on the chart is color coded by the size of the detected anomalous region in the corresponding passage and time. As a result, points with higher counts and connected with other points across several time steps are more salient. This design of anomaly chart is



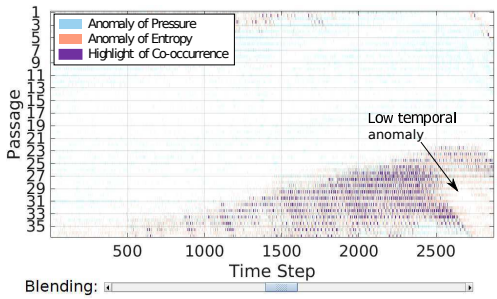
(a) Spatial anomaly plot of pressure.



(b) Spatial anomaly plot of entropy.



(c) Combined spatial anomaly plot of pressure and entropy.



(d) Combined temporal anomaly plot of pressure and entropy.

Figure 5: Spatial and temporal anomaly pattern study for simulation run with CMF = 14.20 kg/s. (c): Showing spatial anomaly of pressure and entropy where the co-occurrence regions are highlighted in blended purple color. (d): Highlighting the co-occurrence of temporal anomaly.

similar to that proposed by Chen et al. [10], but here the anomalies are detected based on the distances of distributions, as discussed in Section 6.

Superimposition. In addition to visualizing the anomaly pattern of a single variable, the expert is interested in how the anomalies detected from different variables are correlated and whether the combination of them generates a more confident indication of stall. Therefore, comparing and contrasting anomalies from different variables are required. We provide superimposition views [20] which composite two anomaly charts into one chart with transparency. Compared to juxtaposition views which place visualizations side-by-side, superimposition does not split the user's attention into different parts of the screen, and also makes the comparison intuitive [20]. Since superimposition may cause visual clutter in complex co-occurrence regions, we provide an interaction tool to overcome this problem as discussed below.

Alpha blending and interaction. To superimpose two selected charts, we overlay them with alpha blending, where the blending coefficient α is adjusted by the user. As shown in Figure 5, α is adjusted using a slider in the bottom, where moving the slider to an end shows the anomalies of a single variable. Therefore, the user can move the slider to contrast different anomaly detection results in position with both contents. Since it is hypothesized that the co-occurrence of anomaly from different detection criteria can indicate the existence of stall cells with more confidence, it is required to highlight these regions on the chart. Therefore, we increase the saturation of co-occurrence regions on the chart when the slider is closer to the middle, as shown in Figure 5c. This enhances the focus on co-occurrence regions and keeps the regions of individual occurrence in context.

Spatial rendering. Since the spatial and temporal anomaly analysis technique quantifies the possibility of containing an anomalous region for the whole data domain, a new scalar field using the anomaly values is constructed. For detailed exploration on the detected anomalous regions in the spatial domain, we allow the expert to render the anomaly field using isosurfaces. The user can either inspect the detected anomalous regions per time step or animate through time to observe the growth of the anomalies. By selecting highly anomalous regions on the anomaly chart, the expert can investigate the location of potential stall impacted regions in data domain and then verify its correlation to the stall inception.

7.2 Uncertain Isocontour Visualization using GMM Data for Hypotheses Verification

In order to help the expert to verify the detected anomalous regions as potential stall cells and further understand the data, it is necessary to provide a flexible exploration tool for spatial data visualization which can utilize the distribution type data. It has been previously shown that with the data represented in the form of spatial local distributions (block-wise GMMs in our case), probabilistic visualization algorithms can be employed for analyzing and visualizing the data with uncertainty quantification [30, 33, 34, 35]. As isocontouring is commonly used by the domain expert to visualize the data variables and verify the stall phenomenon with domain knowledge, we make use of an existing uncertain isocontouring algorithm proposed by Pöthkow et al. [32, 33] to generate the level crossing probability field. For each cubic cell with probability distributions modeled at the eight vertices $(X_1, X_2, \cdots X_8)$, the level-crossing probability of isovalue ϑ is defined as:

$$\begin{aligned} Pr(\vartheta\text{-crossing}) &= 1 - Pr(\vartheta\text{-non-crossing}) \\ &= 1 - Pr(X_1 > \vartheta, X_2 > \vartheta, \cdots X_8 > \vartheta) \\ &\quad - Pr(X_1 < \vartheta, X_2 < \vartheta, \cdots X_8 < \vartheta) \end{aligned} \tag{6}$$

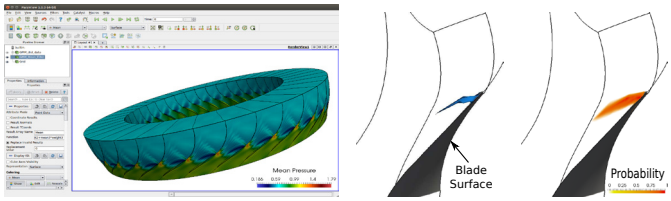
We use the stored GMMs to represent the probability distributions on the vertices of each cell and assign them to X_i. The computation result is a probability field of level crossing, which is then visualized by volume rendering. Figure 6b shows a distribution mean isosurface of a low pressure value (pressure = 0.42) and in Figure 6c the uncertain isocontour of same pressure value is displayed. It can be observed that the uncertain isosurface is able to provide a better estimation with uncertainty information presented in the form of level crossing probability. In this example we have extracted a single passage and applied the aforementioned algorithm for computing the level crossing probability given isovalue of pressure = 0.42.

8 Implementation Details

The proposed approach of *in situ* data summarization followed by a post-processing on the reduced data for the analysis of scientific problems has gained increased attention in the recent past [1, 13].

***In situ* Integration.** TURBO is a complex CFD simulation developed in FORTRAN programming language. The *in situ* analysis code is developed in C++ and linked with the FORTRAN code base in a modular fashion for performing the *in situ* calls. For estimating the initial parameters of the GMMs at the first time step the EM algorithm, OpenCV [6] library was used. Both the domain scientist and the data analysts have worked closely during the integration phase. The *in situ* integration is designed in such a way that the analysis code will be able to directly query the simulation memory for data access. Such a *direct access* scheme is desired since it does not require any deep copy of data. The *in situ* processing and the simulation code in our implementation share the same compute resources, therefore they run in a *synchronous mode*. Before the simulation is started, the expert can select the frequency of *in situ* processing calls and the variables required for post analysis. Also, we have the capability to extract isosurfaces *in situ* if the user already knows what isosurfaces to look for. However, for exploratory analysis tasks like ours, it is difficult to preselect isovalues. Therefore, to allow flexible post-processing we have chosen our output type to be *explorable*, i.e. a wide range of analysis and visualization using such output data will be possible.

VTK/Paraview Integration for GMM based Data Exploration. In order to carry out flexibility in post-hoc data analysis and exploration, the VTK [39] data format is used to store the distributions, which provides extendable self-descriptive data array containers with good data compression. In our work, each of the *in situ* process stores the GMM distribution parameters in VTK arrays and associates them



(a) An example of our GMM distribution mean plugin for Paraview, used to visualize the mean surface of pressure.

(b) Mean isocontour of pressure = 0.42.

(c) Uncertain isocontour of pressure = 0.42.

Figure 6: Visualization of GMM based data using surface renderings and uncertain isocontours.

with structured-grid data points, where each point stores the local distribution of a region from the simulation output. Furthermore, the VTK format is ready to use by the well-known Paraview [3] visualization application. To facilitate probabilistic analysis and uncertainty visualization for the distribution data, we enhanced Paraview and developed plugins for the users to explore the output data from our *in situ* processes. The plugins include filters that generate fields of distribution mean and standard deviation for each data point. This provides an approximated overview of the original data with quantified uncertainty. Figure 6a shows an example of our plugins running on Paraview. The mean of pressure using GMM distributions is computed and visualized on a surface of the rotor. Using the integrated visualization tools provided by Paraview and our added plugins, the expert can easily explore their desired regions of interest using our *in situ* reduced explorable GMM distribution based data with minimal effort.

9 Stall Analysis Results and Expert Feedback

In this section, we present the results for demonstrating the efficacy of our *in situ* distribution guided stall analysis method and discuss the domain expert's feedback. During the development of the *in situ* distribution guided pathway for rotating stall analysis, feedback from the expert was collected regularly which helped us to improve our method to make it more useful.

Experimental Setup. we used two simulation runs in this work to verify the effectiveness of the method. The simulation parameter *corrected mass flow rate* (CMF) was varied to produce two distinct cases where one led to stall and the other without stall. The simulation with CMF = 14.2 kg/s produced stall and the run with CMF = 16.0 kg/s was stable. We verified our method on both the cases and found that the proposed method worked accurately in each of the test cases in detecting stall or the absence of it. Using 4 Gaussians per GMM for estimating the distributions of a data block gave us stable results as was observed in earlier works [15, 27]. Furthermore, since we aimed for capturing local data properties, a smaller block-size was preferable in our case to achieve better accuracy as was suggested by Dutta et al. [15]. However, a smaller block-size could lead to higher storage and hence there should be a trade-off between allowed storage and the block-size. We used a block-size of $5 \times 5 \times 5$ throughout all our experiments which obtained stable results with good data reduction.

9.1 Simulation Run with Stall (CMF = 14.2 kg/s)

Figure 9a shows the mass flow rate plot of this simulation. We ran the simulation for 8 revolutions to obtain a fully developed stall condition. Each revolution consisted of 3600 iterations, hence, a total of 28800 iterations were simulated for this case. The *in situ* call was made at every 10^{th} time step to estimate the GMMs of pressure and entropy variable. Note that the time step numbers used here are in the units of tenths of simulation iterations due to the sampling rate of *in situ* calls.

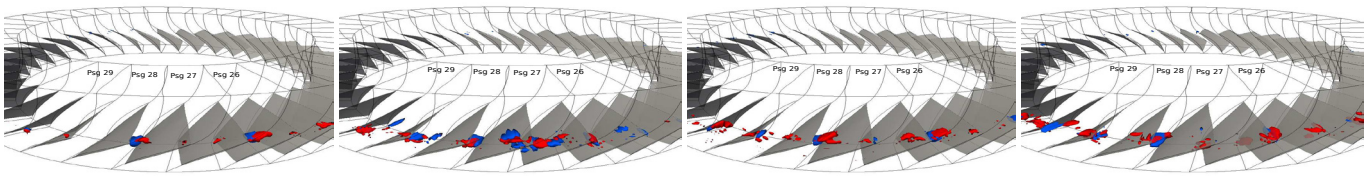
Exploration using spatiotemporal anomaly chart. In Figure 5 we depict the spatial and temporal anomaly charts of this simulation. Figure 5c and 5d show the superimposed chart of spatial and temporal anomaly respectively where each of the chart demonstrates the evolution of anomalous regions by combining both pressure and entropy. From Figure 5c it is observed that around time step 2540, the anomalous regions show strong co-occurrence, and become persistent. This pattern is visible from the consistent purple color. By inspecting the mass flow rate in Figure 9a, we observe that the mass flow rate drops rapidly around the same time step 2540. This sudden drop of mass flow rate confirms the occurrence of stall and verifies that our combined spatial anomaly chart is able to capture this phenomenon. Another important observation from Figures 5c and 5d is that both these combined anomaly charts show the existence of anomalous regions in the passage range 24 - 32 starting around time step 500 which is much earlier than the final occurrence of stall at time step 2540 detected by traditional stall indicator mass flow rate. Furthermore, since the mass flow rate, as shown in Figure 9a, presents an almost flat pattern up to time step 2540, which does not indicate any imminent stall, the expert agreed that the proposed method is able to detect the signs of stall much earlier than the traditional technique using mass flow rate.

By studying the spatial anomaly chart of pressure (Figure 5a), the expert found that there are two anomaly bands. However, entropy anomaly chart showed only one band (Figure 5b) which is located at the bottom of the chart including passages from 24 - 32. Since only the passages in this second band eventually led to stall, the expert concluded that entropy identified the stall impacted regions more accurately than pressure. The expert noted entropy was not a commonly used variable in stall detection and mentioned that with this new finding a more accurate and refined stall indicator measure could be devised using entropy along with pressure.

A further inspection of Figure 5c showed that when the persistent purple regions appear from time step around 2540 reflecting a strong co-occurrence of spatial anomaly of both both pressure and entropy, as annotated in Figure 5c, the temporal anomaly becomes low in such stalled regions as marked in Figure 5d. This pattern of temporal anomaly chart helped the expert to confirm the hypothesis that when the rotating stall is fully developed, the variation of values (pressure and entropy in our study) inside the stall affected regions become less since value distributions of variables do not change with time in fully developed stall cells. Therefore, temporal anomaly is low between the GMMs over time. Furthermore, since such developed stall impacted regions only cover a subset of passages, spatial anomaly becomes high due to the increased asymmetry among blade passages. However, since the expert is primarily interested in detecting signs of stall inception at earlier time steps, it was concluded that both spatial and temporal anomaly methods are capable of capturing the early signs of rotating stall.

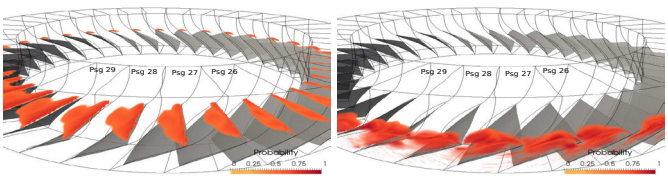
Visualization of detected anomalous locations in spatial domain. To study the anomalous regions detected within the blade passage range 24 - 32, we render the surfaces which contain the detected anomalies of both pressure and entropy. Figure 7 depicts the spatial and temporal anomalous regions of pressure (blue) and entropy (red) detected by the proposed method. Figure 7c and 7d present the anomalous regions at time step 2540 when the sharp drop of mass flow rate is initiated. To investigate the anomalous regions at an earlier time step, results of spatial and temporal anomalous regions from an earlier time step 2200 is shown in 7a and 7b. Two important observations about the detected regions emerged from Figure 7: (1) the anomalous regions of pressure and entropy demonstrate high spatial co-occurrence within the blade passage range 24 - 32 which is also visible from the anomaly chart in Figure 5c and 5d, and (2) the detected anomalous regions appear near the blade tips.

Verification and Expert Feedback. According to the expert, the stall cells are generally located around the blade tip regions. The detected anomalous regions also appear on the tips as seen in Figures 7a-7d. The expert further explains that in a stable state, the tip regions contain a vortex, known as *tip clearance vortex*, and pressure in the vortex core is low. Due to the axisymmetry property, all the tips are expected to have similar low pressure region indicating the tip vortex. However, as the compressor approaches stall, the passages affected by the formation of stall cells tend to show more fluctuations of pressure values around the tip region. Finally when stall occurs the axisymmetry observed in the tip vortices is broken and the region is



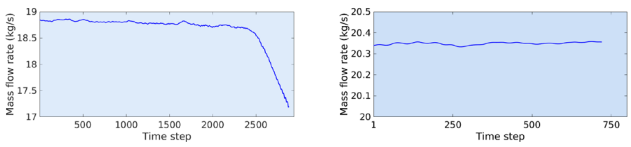
(a) Spatial anomalies at time step 2200. (b) Temporal anomalies at time step 2200. (c) Spatial anomalies at time step 2540. (d) Temporal anomalies at time step 2540.

Figure 7: Visualization of detected anomalous regions with the stall condition (CMF=14.2). Spatial and temporal anomalous regions of pressure (in blue surfaces) and entropy (in red surfaces) are detected near the blade tip regions of several rotor passages. These regions act as blockage to the regular airflow and create flow instability which eventually leads to stall.



(a) Uncertain isocontour of pressure = 0.38. (b) Uncertain isocontour of entropy = 1.0.

Figure 8: Uncertain isocontour visualization at time step 2540 for visual exploration and verification of stall impacted regions.

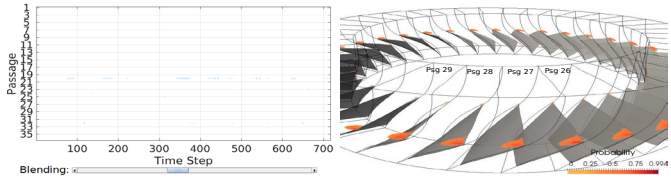


(a) Simulation with CMF = 14.2 kg/s (b) Simulation with CMF = 16.0 kg/s

Figure 9: The mass flow rate plot of simulations in a stall condition (CMF = 14.2 kg/s) and a stable condition (CMF = 16.0 kg/s).

classified as anomalous. Entropy values in the stall impacted regions also increases significantly compared to the blade passages which do not contain the stall cells. To visualize these phenomena, the expert used uncertain isocontours of low pressure (pressure=0.38) and high entropy (entropy=1.0) as shown in Figure 8. As can be seen from Figure 8a, the uncertain isocontour of pressure at time step 2540 is distinctly different in the stall affected passages, while on the other side of the rotor, the contours are well organized and symmetric. Also, in Figure 8b it is observed that high entropy contour is located in the similar passages close to tip regions which further confirms the locations of stall. These observations conform well with the blade passages detected using our anomaly based analysis and validate the efficacy of the proposed method.

To further confirm whether the anomalous regions are indeed stall cells, the expert studied the formation and evolution of anomalous regions for all the time steps. It was observed that the anomalous regions actually propagate from passage to passage in the opposite direction to the rotation of blades. During the formation of these regions, they pop up and gradually move to the neighboring passages. This behavior is consistent to the transportation of stall cells and a domain explanation is as follows: when a stall cell grows in a passage, it forms a blockage to the incoming flow which redirects a portion of the flow to the neighboring blades. This increases the angle of attack and causes stall for the proceeding blade, as well as, decrease the angle of attack on the preceding blade increasing stability. However, as the proceeding blade stalls, the currently stalled blade experiences a decrease in the angle of attack and begins to resume normal operation. The cycle of stall cell passing continues and thereby causes the counter-rotating motion observed in the simulation. With these explanations, the expert finally concluded that the detected anomalous regions are stall cells.



(a) Combined spatial anomaly plot. Almost no anomalous regions are detected. (b) Uncertain isocontour of pressure = 0.38. Well structured symmetric isocontour of pressure is observed.

Figure 10: Anomaly analysis and spatial visualization of the stable condition (CMF = 16.0 kg/s).

9.2 Simulation Run without Stall (CMF = 16.0 kg/s)

The simulation run with CMF = 16.0 kg/s is a known configuration where the simulation runs consistently and is considered to be stable. It demonstrates high axisymmetry across all the passages. For verification, we ran this configuration for 2 revolutions i.e. 7200 time steps and the *in situ* call was made at every 10th time step. In Figure 9b we show the mass flow rate plot of this simulation and observe a flat trend. As can be seen in Figure 10a, the combined spatial anomaly chart of pressure and entropy barely detects any anomalous regions and hence the chart is almost clean. This confirms the effectiveness of the proposed distribution based anomaly analysis methods in differentiating a stable and unstable operating condition. To verify the data in spatial domain, in Figure 10b the uncertain isocontour of pressure = 0.38 is depicted. From Figure 10b it is observed that all the passages have similar pressure value distributions and hence they produce similar isocontours following the axisymmetry property.

9.3 Discussion of the Stall Analysis Results

The above results on different parameter conditions demonstrate the capability of the proposed *in situ* distribution guided approach for rotating stall analysis. Our technique shows the benefits of distribution based analysis when the target feature, i.e. the stall cell, has no precise descriptor. Furthermore, the local region based spatial and temporal anomaly analysis also demonstrates the efficacy of our method in detecting earlier signs of stall as depicted in anomaly charts in Figure 5, whereas the traditional approach using mass flow rate does not work well. From our comparative visualization of anomaly plots, the expert also discovered that the entropy anomaly indicates the potential stall impacted regions more accurately than pressure anomaly alone. He concluded that by finding the co-occurrence of both pressure and entropy anomalies, a more refined stall detection technique can be obtained. Another hypothesis of the expert that the variation of data values inside a fully developed stall cell becomes less compared to the stall inception phase is also confirmed by analyzing the temporal anomaly chart where the degree of measured temporal anomaly diminishes as marked in Figure 5d. Finally, by rendering uncertain isocontours the expert is able to visualize the data properties in spatial domain and validate the results of the proposed approach. Next, we provide a quantitative performance study of the proposed *in situ* approach and show that the method achieves significant savings in both storage and computation time.

Table 1: Post-hoc GMM computation time with I/O in the absence of *in situ* processing.

Component	2 revs.	4 revs.	8 revs.
Simulation raw I/O (hrs)	2.59	5.2	10.36
GMM computation (hrs)	2.38	4.82	9.52

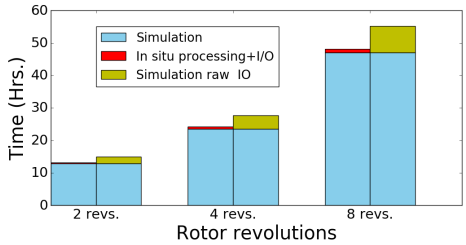


Figure 11: Timing comparison with and without raw output. With the *in situ* pathway, the raw I/O time can be saved.

10 Performance Study

The performance study was done using a cluster, Oakley [8], at the Ohio Supercomputer Center, which contains 694 nodes with Intel Xeon x5650 CPUs (12 cores per node) and 48 GB of memory per node. A parallel high-performance, and shared disk space Lustre was used for I/O during the simulation runs with *in situ* processing.

Storage savings. A full annulus run of TURBO with 1 rotor revolution generates 5.04 TBs of raw data. In our two test cases, we ran the simulation for 8 revolutions for the first test case with CMF = 14.2 to capture the stall phenomenon and 2 revolutions for the second case with CMF = 16.0. These two runs generated raw data of 40.32 TBs and 10.08 TBs respectively. The *in situ* call was made at every 10^{th} time step which required us to process 4.032 TBs for the first and 1.008 TBs for the second test case. The simulation model has three sections: one rotor and 2 stators. In these experiments, we have stored GMMs only for the rotor which is the focused region of study, and have also stored only 2 variables. The data size for the rotor part in plot3d format is 690 MB per time step. The output of the *in situ* summarized data of two variables, in VTK multi-block format, took only 51.8 GBs for the first simulation run and 12.9 GBs for the second run, which is significantly less than the actual raw data size needed for a purely post-hoc analysis. An important point to mention is that with a different CMF condition, we would require to run the simulation for longer time and the size of raw data would be even larger.

Computation time savings. In Figure 11, we present the comparison of timings for the two scenarios: with and without *in situ* processing. The left bar in each case shows the simulation time (light blue) along with the *in situ* processing time (red), and the right bar shows simulation time (light blue) with the raw output time (green). Note that the *in situ* processing timings include the I/O time for GMM distributions, which is significantly less than the actual raw data output time. Without the proposed *in situ* processing (i.e. the GMM computation and distribution type data I/O), in addition to the mandatory raw data I/O time, extra time for estimating the GMMs post-hoc using the raw data is necessary. This extra time without the *in situ* scenario is presented in Table 1, where the I/O and computation time become prohibitive as the data size grows with increased rotor revolutions.

In order to study the overhead of *in situ* processing, we tested our approach using a half annulus model of TURBO which consists of 18 blade passages in stead of 36. In this half annulus configuration, the workload for each processor was kept the same as full annulus. In Table 2, we report the percentage timings of both half and full annulus *in situ* runs. From Table 2, we observe that the percentage time required for our *in situ* processing is only a small fraction, around 2.5% of the simulation time in both the cases. Therefore, the benefits obtained in terms of saving time in post-hoc exploration using the proposed *in situ* strategy is obvious, since we essentially bypass the simulation raw I/O, post-hoc GMM computation and I/O time completely by performing

Table 2: Percentage timing of *in situ* processing with half and full annulus runs. All the cases show similar percentage.

Configuration	2 revs.		4 revs.	
	Simulation	In situ	Simulation	In situ
Half annl. (164 cores)	97.3%	2.7%	97.5%	2.5%
Full annl. (328 cores)	97.63%	2.37%	97.42%	2.58%

Table 3: Computation time including I/O for anomaly analysis.

Anomaly type	2 revs.	4 revs.	8 revs.
Temporal anomaly analysis time (hrs)	0.56	1.11	2.19
Spatial anomaly analysis time (hrs)	0.57	1.12	2.20

the task *in situ*. The timings of spatial and temporal anomaly analysis using the reduced GMM distribution data are shown in Table 3 which include the I/O time as well. Hence, by performing *in situ* processing, we have enabled a scalable and flexible post-hoc rotating stall analysis to help the expert achieve a better understanding of the phenomenon.

11 Conclusions

In this work, we have demonstrated the effectiveness of a distribution guided, local region based rotating stall analysis. The approach that takes advantage of *in situ* processing for summarizing the important data in simulation time. Our method uses mixtures of Gaussians which facilitates flexible and scalable post-hoc analysis. By exploiting the spatiotemporal variations of distributions, statistically anomalous regions in the data are identified which have been shown to have strong correlation to the inception of rotating stall. The performance section also shows that by following the *in situ* pathway, significant cost reduction is achieved in terms of both storage and post-hoc computation. In the future, we would like to enable the user to steer the simulation by changing the CMF parameter and provide real-time feedback of stall analysis results to the expert. Furthermore, we would like to extend our work to include more sophisticated *in situ* uncertainty quantification capabilities and apply it on other parameter configurations.

Acknowledgments

This work was supported in part by NSF grants IIS- 1250752, IIS-1065025, and US Department of Energy grants DE- SC0007444, DE-DC0012495, program manager Lucy Nowell.

References

[1] J. Ahrens, S. Jourdain, P. OLeary, J. Patchett, D. H. Rogers, and M. Petersen. An image-based approach to extreme scale in situ visualization and analysis. In *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 424–434, 2014.

[2] T. Athawale, E. Sakhaee, and A. Entezari. Isosurface visualization of data with nonparametric models for uncertainty. *IEEE Trans. on Vis. and Comp. Graphics*, 22(1):777–786, 2016.

[3] U. Ayachit. *The ParaView Guide: A Parallel Visualization Application*. Kitware Inc., 4.3 edition, 2015. ISBN 978-1-930934-30-6.

[4] J. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, 1998.

[5] G.-P. Bonneau, H.-C. Hege, C. R. Johnson, M. M. Oliveira, K. Potter, P. Rheingans, and T. Schultz. *Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization*, chapter Overview and State-of-the-Art of Uncertainty Visualization, pages 3–27. Springer London, 2014.

[6] G. Bradski. The OpenCV Library. *Dr. Dobb's J. of Software Tools*, 2000.

[7] K. Brodlie, R. Allendes Osorio, and A. Lopes. *Expanding the Frontiers of Visual Analytics and Visualization*, chapter A Review of Uncertainty in Data Visualization, pages 81–109. Springer London, 2012.

[8] O. S. Center. Oakley supercomputer. `http://osc.edu/ark:/19495/hpc0cvqn`, 2012.

[9] A. Chaudhuri, T.-H. Wei, T.-Y. Lee, H.-W. Shen, and T. Peterka. Efficient range distribution query for visualizing scientific data. In *IEEE Pacific Visualization Symposium (PacificVis), 2014*, pages 201–208, 2014.

[10] C.-M. Chen, S. Dutta, X. Liu, G. Heinlein, H.-W. Shen, and J.-P. Chen. Visualization and analysis of rotating stall for transonic jet engine simulation. *IEEE Trans. on Vis. and Comp. Graphics*, 22(1):847–856, 2016.

[11] J. Chen, R. Webster, M. Hathaway, G. Herrick, and G. Skoch. Numerical simulation of stall and stall control in axial and radial compressors. In *44th AIAA Aerospace Sciences Meeting and Exhibit*. American Institute of Aeronautics and Astronautics, 2006.

[12] J.-P. Chen, M. D. Hathaway, and G. P. Herrick. Prestall behavior of a transonic axial compressor stage via time-accurate numerical simulation. *Journal of Turbomachinery*, 130(4):041014, 2008.

[13] H. Childs. Data exploration at the exascale. *Supercomputing frontiers and innovations*, 2(3), 2015.

[14] I. J. Day, T. Breuer, J. Escuret, M. AU - Cherrett, and A. AU - Wilson. Stall inception and the prospects for active control in four high-speed compressors. *Journal of Turbomachinery*, 121(1):18–27, 1999.

[15] S. Dutta and H.-W. Shen. Distribution driven extraction and tracking of features for time-varying data analysis. *IEEE Trans. on Vis. and Comp. Graphics*, 22(1):837–846, 2016.

[16] N. Fabian, K. Moreland, D. Thompson, A. C. Bauer, P. Marion, B. Gevecik, M. Rasquin, and K. E. Jansen. The paraview coprocessing library: A scalable, general purpose in situ visualization library. In *2011 IEEE Symposium on Large Data Analysis and Visualization (LDAV)*, pages 89–96, 2011.

[17] R. Haimes. pv3: A distributed system for large-scale unsteady cfd visualization. In *AIAA paper*, pages 94–0321, 1994.

[18] M. Hathaway, G. Herrick, J. Chen, and R. Webster. Time accurate unsteady simulation of the stall inception process in the compression system of a US army helicopter gas turbine engine. In *31st Annual International Symposium on Computer Architecture, 2004. Proceedings*, pages 166–177, 2004.

[19] H. Jänicke, A. Wiebel, G. Scheuermann, and W. Kollmann. Multifield visualization using local statistical complexity. *IEEE Trans. on Vis. and Comp. Graphics*, 13(6):1384–1391, 2007.

[20] W. Javed and N. Elmqvist. Exploring the design space of composite visualization. In *2012 IEEE Pacific Visualization Symposium (PacificVis)*, pages 1–8, 2012.

[21] C. Johnson and J. Huang. Distribution-driven visualization of volume data. *IEEE Trans. on Vis. and Comp. Graphics*, 15(5):734–746, 2009.

[22] D. Kao, A. Luo, J. L. Dungan, and A. Pang. Visualizing spatially varying distribution data. In *Proceedings of the Sixth International Conference on Information Visualisation, 2002*, pages 219–225, 2002.

[23] V. Karavasilis, C. Nikou, and A. Likas. Visual tracking using the earth mover's distance between gaussian mixtures and kalman filtering. *Image and Vision Computing*, 29(5):295–305, 2011.

[24] T.-Y. Lee and H.-W. Shen. Efficient local statistical analysis via integral histograms with discrete wavelet transform. *IEEE Trans. on Vis. and Comp. Graphics*, 19(12):2693–702, 2013.

[25] H. Lehmann and B. Jung. In-situ multi-resolution and temporal data compression for visual exploration of large-scale scientific simulations. In *IEEE 4th Symposium on Large Data Analysis and Visualization (LDAV), 2014*, pages 51–58, 2014.

[26] H. Ling and K. Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(5):840–853, 2007.

[27] S. Liu, J. Levine, P. Bremer, and V. Pascucci. Gaussian mixture model based volume visualization. In *2012 IEEE Symposium on Large Data Analysis and Visualization (LDAV)*, pages 73–77, 2012.

[28] J. F. Lofstead, S. Klasky, K. Schwan, N. Podhorszki, and C. Jin. Flexible IO and Integration for Scientific Codes Through the Adaptable IO System (ADIOS). In *Proceedings of the 6th International Workshop on Challenges of Large Applications in Distributed Environments*, CLADE '08, pages 15–24. ACM, 2008.

[29] C. Lundstrom, P. Ljung, and A. Ynnerman. Local histograms for design of transfer functions in direct volume rendering. *IEEE Trans. on Vis. and Comp. Graphics*, 12(6):1570–1579, 2006.

[30] A. Luo, D. Kao, and A. Pang. Visualizing spatial distribution data sets. In *Proceedings of the Symposium on Data Visualisation 2003*, VISSYM '03, pages 29–38, 2003.

[31] N. M. McDougall, N. A. Cumpsty, and T. P. Hynes. Stall inception in axial compressors. *Journal of Turbomachinery*, 112(1):116–123, 1990.

[32] K. Pöthkow and H.-C. Hege. Positional uncertainty of isocontours: Condition analysis and probabilistic measures. *IEEE Trans. on Vis. and Comp. Graphics*, 17:1393–1406, 2011.

[33] K. Pöthkow, B. Weber, and H.-C. Hege. Probabilistic marching cubes. In *Proceedings of the 13th Eurographics / IEEE - VGTC Conference on Visualization*, EuroVis'11, pages 931–940, 2011.

[34] K. Potter, J. Kniss, R. Riesenfeld, and C. R. Johnson. Visualizing summary statistics and uncertainty. *Computer Graphics Forum (Proceedings of Eurovis 2010)*, 29(3):823–831, 2010.

[35] K. Potter, J. Krüger, and C. Johnson. Towards the visualization of multidimensional stochastic distribution data. In *Proceedings of The International Conference on Computer Graphics and Visualization (IADIS) 2008*, 2008.

[36] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[37] B. E. Ruttenberg and A. K. Singh. Indexing the earth mover's distance using normal distributions. *Proceedings of the VLDB Endowment*, 5(3):205–216, 2011.

[38] R. Sandler and M. Lindenbaum. Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE trans. on pattern analysis and machine intelligence*, 33(8):1590–1602, 2011.

[39] W. Schroeder, K. Martin, and B. Lorensen. *The Visualization Toolkit: An Object Oriented Approach to 3D Graphics*. Kitware Inc., fourth edition, 2004. ISBN 1-930934-19-X.

[40] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1999*, volume 2, page 252 Vol. 2, 1999.

[41] D. Thompson, J. A. Levine, J. C. Bennett, P. T. Bremer, A. Gyulassy, V. Pascucci, and P. P. Pbay. Analysis of large-scale scalar data using hixels. In *Large Data Analysis and Visualization (LDAV), 2011 IEEE Symposium on*, pages 23–30, 2011.

[42] V. Vishwanath, M. Hereld, and M. E. Papka. Toward simulation-time data analysis and i/o acceleration on leadership-class systems. In *2011 IEEE Symposium on Large Data Analysis and Visualization (LDAV)*, pages 9–14, 2011.

[43] C. Wang, H. Yu, and K.-L. Ma. Importance-driven time-varying data visualization. *IEEE Trans. on Vis. and Comp. Graphics*, 14(6):1547–1554, 2008.

[44] C. Wang, H. Yu, and K. L. Ma. Application-driven compression for visualizing large-scale time-varying data. *IEEE Computer Graphics and Applications*, 30(1):59–69, 2010.

[45] Y. Wang, W. Chen, J. Zhang, T. Dong, G. Shan, and X. Chi. Efficient volume exploration using the gaussian mixture model. *IEEE Trans. on Vis. and Comp. Graphics*, 17(11):1560–1573, 2011.

[46] T.-H. Wei, C.-M. Chen, and A. Biswas. Efficient local histogram searching via bitmap indexing. *Computer Graphics Forum*, 34(3):81–90, 2015.

[47] M. Werman, S. Peleg, and A. Rosenfeld. A distance metric for multidimensional histogram. *CVGIP: Graphical Models and Image Processing*, 32(3):328–336, 1983.

[48] B. Whitlock, J. M. Favre, and J. S. Meredith. Parallel in situ coupling of simulation with a fully featured visualization system. In *Proceedings of the 11th Eurographics Conference on Parallel Graphics and Visualization*, EGPGV '11, pages 101–109. Eurographics Association, 2011.

[49] J. Woodring, J. Ahrens, J. Figg, J. Wendelberger, S. Habib, and K. Heitmann. In-situ sampling of a large-scale particle simulation for interactive visualization and analysis. In *Proceedings of the 13th Eurographics / IEEE - VGTC Conference on Visualization*, pages 1151–1160. Eurographics Association, 2011.

[50] J. Woodring, J. Ahrens, T. J. Tautges, T. Peterka, V. Vishwanath, and B. Geveci. On-demand unstructured mesh translation for reducing memory pressure during in situ analysis. In *Proceedings of the 8th International Workshop on Ultrascale Visualization*, pages 3:1–3:8. ACM, 2013.

[51] J. Woodring, M. Petersen, A. Schmeier, J. Patchett, J. Ahrens, and H. Hagen. In situ eddy analysis in a high-resolution ocean climate model. *IEEE Trans. on Vis. and Comp. Graphics*, 22(1):857–866, 2016.

[52] H. Yu, C. Wang, R. W. Grout, J. H. Chen, and K. L. Ma. In situ visualization for large-scale combustion simulations. *IEEE Computer Graphics and Applications*, 30(3):45–57, 2010.