Data-Driven Estimation of Temporal-Sampling Errors in Unsteady Flows

Harsh Bhatia, Steve N Petruzza, Rushil Anirudh, Attila G Gyulassy, Robert M Kirby, Valerio Pascucci and Peer-Timo Bremer

October 19, 2021

Data-Driven Estimation of Temporal-Sampling Errors in Unsteady Flows

Harsh Bhatia[1], Steve N. Petruzza[2], Rushil Anirudh[1], Attila G. Gyulassy[3],
Robert M. Kirby[3], Valerio Pascucci[3], and Peer-Timo Bremer[1]

[1] Center for Applied Scientific Computing, Lawrence Livermore National Laboratory,
Livermore, CA 94551
{hbhatia, anirudh1, ptbremer}@llnl.gov
[2] Department of Computer Science, Utah State University, Logan, UT 84322
steve.petruzza@usu.edu
[3] Scientific Computing & Imaging Institute, University of Utah, Salt Lake City, UT
84112 {kirby, jediati, pascucci}@sci.utah.edu

Abstract. While computer simulations typically store data at the highest available spatial resolution, it is often infeasible to do so for the temporal dimension. Instead, the common practice is to store data at regular intervals, the frequency of which is strictly limited by the available storage and I/O bandwidth. However, this manner of temporal subsampling can cause significant errors in subsequent analysis steps. More importantly, since the intermediate data is lost, there is no direct way of measuring this error after the fact. One particularly important use case that is affected is the analysis of unsteady flows using pathlines, as it depends on an accurate interpolation across time. Although the potential problem with temporal undersampling is widely acknowledged, there currently does not exist a practical way to estimate the potential impact. This paper presents a simple-to-implement yet powerful technique to estimate the error in pathlines due to temporal subsampling. Given an unsteady flow, we compute pathlines at the given temporal resolution as well as subsamples thereof. We then compute the error induced due to various levels of subsampling and use it to estimate the error between the given resolution and the unknown ground truth. Using two turbulent flows, we demonstrate that our approach, for the first time, provides an accurate, *a posteriori* error estimate for pathline computations. This estimate will enable scientists to better understand the uncertainties involved in pathline-based analysis techniques and can lead to new uncertainty visualization approaches using the predicted errors.

Keywords: Unsteady flow · Sampling errors · Temporal resolution · Uncertainty visualization

1 Introduction

Unsteady flows describe many natural and artificial phenomena and form the core of a large number of science and engineering applications [6,23,34]. In many cases, the primary focus is on understanding the transport of material in the flow, typically represented using *pathlines* — paths of massless particles advected by the time-varying flow (see Equation 1). However, as pathlines are typically computed using iterative, numerical integration, they are susceptible to errors due to a number of sources.

In practice, errors due to insufficient temporal sampling of data are often considered to be most challenging for two main reasons: (1) the lack of data can be severe, and (2) the corresponding error cannot be easily computed. Virtually no large-scale simulation can afford to also store all available time-steps as this would increase simulation time by orders of magnitude and create unmanageable amounts of data. Instead, the data is subsampled in time, and often only every 500th or 1000th snapshot is available for analysis [12]. Since all intermediate data is lost, the error resulting from the subsampling cannot be directly computed and is often accepted as an inevitable consequence of the storage and I/O limitations. However, especially for the large-scale, turbulent simulations of greatest interest, the unknown error may dramatically impact computed pathlines.

Motivating Case Study. We consider a large-scale combustion simulation of a lifted jet flame [34,35] performed using S3D [13]. Such flows are used to study direct-injection spark ignition engines for commercial boilers as well as fundamental combustion phenomena. The simulation uses a $2025 \times 1600 \times 400$ rectilinear grid and captures several observables, such as velocities and temperature, resulting in about 280 GB of data per time-step. S3D uses an explicit Runge-Kutta (RK) integration scheme with a step size of 4×10^{-9} units. However, due to the to large I/O overheads and storage limitations, only every 500th snapshot is stored. It is important to note that the scientists consider this temporal resolution, *i.e.*, 2×10^{-6} units, *exceptionally high* for this type of study.

What make this simulation of particular interest for this paper is that it also includes a set of tracer particles computed *in situ*, which offers an opportunity to study errors introduced through temporal subsampling. Specifically, a total of 54,935 particles were tagged and traced alongside the simulation and stored at a step size of 2×10^{-7}. Particles are available for a total of 299 time-steps uniformly distributed in the time-range $[1.7, 1.7598] \times 10^{-3}$, effectively defining a set of highly-accurate pathlines. To compare these *in situ* pathlines with the ones computed in post-processing, we consider pathlines computed from the saved data covering the same range with the identical starting position. Since the data is spatially over-resolved, we use trilinear interpolation in space and the traditional linear interpolation in time with a conservative step-size of 2×10^{-8}.

Fig. 1 provides a visual comparison between the *in situ* (particles) and the *post hoc* pathlines to highlight the differences, e.g., how the pathlines on the right fail to capture the clear separation of the flow between the top and bottom layers in the flame. Fig. 2 shows the distribution of point-wise errors between the two sets of pathlines (computed using Equation 2) in spherical coordinates and conveys that most pathlines differ from the corresponding *in situ* particle paths by about 16 grid cells, and a substantial number of pathlines deviate by up to 50 grid cells. Similarly, the distribution of the azimuthal angle, φ,

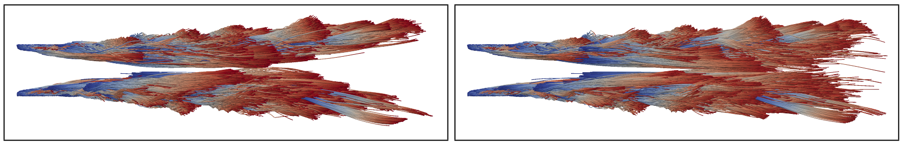

Fig. 1. A visual comparison of *in situ* (left) and *post hoc* (right) pathlines illustrates that the latter can misrepresent flow behavior, as they are affected by temporal subsampling errors. Pathlines are colored blue-to-red on time $[1.7, 1.7598] \times 10^{-3}$.

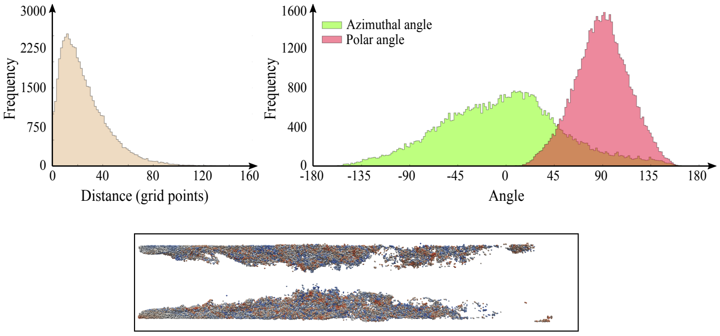

Fig. 2. Top: Distributions of point-wise differences between the two set of pathlines indicate high variance in error. Bottom: Spatial mapping of the polar angles of the difference to the corresponding seed points highlights that the errors appear to be distributed randomly, which may cause substantial artifacts in subsequent analysis.

also has high variance and highlights that the computed pathlines are almost equally likely to be "ahead of" (faster than) the *in situ* particles ($|\varphi| < 90°$) or "behind" (slower than) them ($|\varphi| > 90°$). Fig. 2 also maps the polar angle of the differences to the seed points of corresponding pathlines showing that they appear to be randomly distributed in space. This error behavior is of concern as one of the main uses of pathlines is the computation of derived quantities, such as the finite-time Lyapunov exponent (FTLE) [19]. The FTLE is defined through a spatial derivative of the particle positions and the random errors shown in Fig. 2 would be expected to cause substantial artifacts in the results. Nevertheless, without the *in situ* particles for verification, which only a few simulations provide, it is challenging to determine the expected accuracy of *post hoc* pathlines and conveying the resulting uncertainty.

Given that the sampling rate for this flow is considered exceptionally high, *post hoc* pathlines would likely have been accepted as a reliable approximation of the ground truth. However, the comparatively large and random errors discussed above raise significant concerns on the reliability of any pathline-based analysis. Although one would expect similar problems in other types of large-scale simulations of unsteady flows as well, without an understanding of the extent of inherent errors, scientists must currently choose between disregarding interesting results, because they cannot be validated, or accepting *post hoc* pathlines as best available information, potentially arriving at incorrect conclusions.

Contributions. To address this challenge, we present a data-driven approach to model subsampling errors in pathlines. Our *a posteriori* error estimate provides users with insights into the likely effects of temporal subsampling *without* access to the ground truth data. In particular, we compute pathlines at the given resolution as well as at successively coarser subsampled resolutions. Using two turbulent flows, we show that, in general, the differences between pathlines of successively-coarser resolutions can be modeled using a simple, supervised regression model that allows predicting the differences between the finest available resolution and the unknown ground truth. Our approach requires no additional implementation beyond the existing pathline computation and comparison, yet provides a reliable *a posteriori* error estimate for pathlines.

2 Related Work

Analyzing pathlines of unsteady flows is among the most fundamental ways of understanding its dynamic behavior [24]. *Pathlines* represent the path taken by a massless particle as it is advected by the flow. They have been used to compute the topological segmentation of 2D flows [30]; constructs similar to streak surfaces have been used for the topological analysis of 3D flows [15,22,31]. The notion of pathlines has also been extended to inertial particles to address more realistic physical phenomena [9,18]. Pathlines are often used to compute the FTLE [19] or the finite-space Lyapunov exponent (FSLE) [28], which are defined using the spatial derivative of the positions of neighboring seed particles after a given amount of time or distance, respectively. The FTLE and FSLE are believed to highlight the Lagrangian coherent structures (LCS) in the flow [20], such as material boundaries. However, dependent upon derivatives, the FTLE, the FSLE, and, hence, the LCS are highly sensitive to the errors in pathlines.

To date, the potential problems regarding uncertainty in pathlines remain largely unaddressed, despite their importance in the analysis of unsteady flows. In practice, pathlines are computed through numerical schemes, such as RK integration [7] with a high-order interpolation in space and a linear interpolation in time. It is well known that numerical integration is prone to compounding errors [14], especially if the source data is noisy or under-sampled. Almost all error studies in this context have focused on either the steady (time-independent) case or the analysis of uncertain data or errors in the integration. For example, there exist uncertainty visualization techniques for enhanced glyphs [27,33] to represent fields, and thick tubes [21,32] or streamwaves [3,4] to represent uncertainty in streamlines. Otto *et al.* [26,25] simulate uncertainty in data stochastically, but disregard uncertainty due to the computation of streamlines. For pathlines, Teitzel *et al.* [29] study the error resulting from numerical integration and compare different RK techniques with an additional focus on performance and Darmofal and Haimes [14] provide detailed analysis of different integration schemes. Chen *et al.* [11] addressed the problem of integration uncertainty in sampled data by modeling the errors with Gaussian distributions. Nevertheless, no techniques exist that estimate and visualize subsampling uncertainty in unsteady flows, and, the potential errors from a lack of temporal resolution have largely been ignored.

Recently, new Lagrangian representations have been proposed to alleviate the dependence of conventional representations on temporal resolution. Specifically, Agranovsky *et al.* [1] propose to compute and save a set of basis pathlines *in situ* with high accuracy and use those to compute any pathline in the post-processing. Along the same lines, Chandler *et al.* [10] utilize densely sampled in-situ particles to compute pathlines, thus reducing the numerical integration of pathlines to geometric interpolation. These representations alleviate some of the challenges of low temporal resolution at the cost of new errors when remapping particles between *in situ* pathlines. Nevertheless, assuming a sufficiently dense set of *in situ* pathlines, the remapping errors appear significantly smaller than the errors due to temporal subsampling. Unfortunately, very few simulations will natively compute *in situ* particles, and there exist a number of related challenges, such as, automatically computing good seed points. As a result, the applicability of these ideas is currently limited, and it remains important to find better ways to understand temporal subsampling errors in the current analysis pipeline.

3 Temporal Subsampling Errors in Pathlines

We consider the flow computed at the time-step of the simulation to be the ground truth and denote it as $\vec{V}_1(\mathbf{x}, t)$. Although $\vec{V}_1(\mathbf{x}, t)$ typically contains modeling and simulation errors and, therefore, in principle, may not be "correct" compared to the physical phenomenon under consideration, the analysis cannot be more accurate than the initial simulation itself, making this a reasonable assumption. The question we aim to answer is: *given a (temporally) subsampled flow $\vec{V}_k(\mathbf{x}, t)$, which contains timesteps only at some (temporal) resolution $k > 1$, and a given algorithm to compute pathlines, how much is a pathline computed in $\vec{V}_k(\mathbf{x}, t)$ expected to differ from its counterpart computed for $\vec{V}_1(\mathbf{x}, t)$?*

Computing pathlines and measuring errors. Pathlines represent paths of massless particles advected in the flow, given by the solution of the following integration of an ordinary differential equation.

$$\mathbf{p}(t) = \mathbf{p}_0 + \int_{t_0}^{t} \vec{V}(\mathbf{p}(\tau), \tau)\, \mathrm{d}\tau, \tag{1}$$

with $\mathbf{p}_0 = \mathbf{p}(t_0)$. To explore sampling errors in pathlines, we use standard ways to interpolate flows and compute pathlines, keeping all parameters constant, varying only the temporal resolution. Specifically, we use a RK 4-5 integrator [7] with trilinear interpolation in space and linear interpolation in time.

Given two pathlines $\mathbf{p}(t)$ and $\mathbf{q}(t)$ with the same seed position $\mathbf{p}_0 = \mathbf{q}_0$, but computed at different temporal resolutions, we measure the error between them in terms of their maximal pointwise distances, *i.e.*,

$$\varepsilon(\mathbf{p}, \mathbf{q}) = \max_{0 \le \tau \le t} \|\mathbf{p}(\tau) - \mathbf{q}(\tau)\|. \tag{2}$$

We choose the maximal error as one typically wants to understand the worst case impact on any downstream analysis. Note that since $\mathbf{p}(t)$ and $\mathbf{q}(t)$ are computed using different temporal resolutions, care must be taken that the points corresponding to the same value of time are compared.

3.1 Temporal Subsampling of Simulated Unsteady Flows

Let $\vec{V}_{\Delta}(\mathbf{x}, t)$ represent a flow sampled at temporal resolution $\Delta \geq 1$. As discussed above, $\Delta = 1$ denotes the simulation time-step (the ground truth) and $\Delta = k$ the given sampling rate, *i.e.*, the flow stored at every k^{th} timestep. For a pathline computed at $\Delta = k$, the goal is to estimate the error introduced by temporal subsampling with respect to the ground truth. We denote this error as $\varepsilon_{(k,1)}$.

In order to estimate this error without requiring the ground truth, we further subsample the given data to resolutions $2k$, $3k$, ..., nk, and study the resulting errors between the corresponding pathlines at successive levels of subsampling, *i.e.*, $\varepsilon_{(2k,k)}$, $\varepsilon_{(3k,2k)}$, ..., $\varepsilon_{(nk,(n-1)k)}$. Specifically, we compute pathlines at these resolutions and analyze how coarser resolutions are related to the finer ones. Each successive subsampling is likely to introduce additional errors, and we expect them to be proportional to their magnitude, *i.e.*, pathlines with high $\varepsilon_{(2k,k)}$ are expected to show high $\varepsilon_{(3k,2k)}$. Therefore, we assume that the relationship between errors introduced at every level of subsampling can be modeled as

$$\varepsilon_{((n+1)k,nk)} = m_{nk} \; \varepsilon_{(nk,(n-1)k)}, \tag{3}$$

where m_{nk} is a resolution-dependent constant that quantifies the loss of information between the two subsampling steps.

We note that m_{nk} is not just influenced by the effective resolution nk but also by the "type" of pathline under consideration. A low m_{nk} indicates that the additional subsampling did not cause any significant increase in error. Typically, we expect a low m_{nk} for pathlines that are mostly laminar and, hence, can be accurately computed at lower resolutions, or pathlines, which at nk, already contain such a large error that further subsampling does not have a significant effect. On the other hand, for turbulent pathlines still containing meaningful information, one would expect m_{nk} to change significantly for different n, as a substantial amount of information may be lost at each level of subsampling. Furthermore, for most turbulent flows, we expect the value of m_{nk} to decay with subsampling, as most of the information is lost during the initial subsampling, whereas a relatively-smaller loss of information is incurred at later stages.

A similar technique to estimate sampling errors by upsampling and downsampling in the context of high-definition images was described by Berger *et al.* [2]. Whereas they used a spine-tube interpolant to predict error for spatial subsampling, our goal is to estimate errors due to temporal subsampling.

3.2 Data-Driven Modeling of Errors

Consider the generalization of Equation 3 as

$$\varepsilon_{(k,1)} = f(\varepsilon_{(2k,k)}, \; \varepsilon_{(3k,2k)}, \; \ldots, \; \varepsilon_{(nk,(n-1)k)}), \tag{4}$$

which parameterizes the error $\varepsilon_{(k,1)}$ as a function of errors occuring at lower sampling resolutions. If $f(\cdot)$ is linear, Equation 4 generalizes Equation 3 by including more than a single error with a nonzero weight. Recall that given $n-1$ subsampling errors for a pathline, $\varepsilon_{(2k,k)}$, $\varepsilon_{(3k,2k)}$, $\ldots$, $\varepsilon_{(nk,(n-1)k)}$, the goal is to predict $\varepsilon_{(k,1)}$. However, in most practical cases, the ground truth data is not available to validate our prediction; instead, we validate our model by predicting $\varepsilon_{(2k,k)}$ having observed the errors for subsequent resolutions.

The function $f(\cdot)$ can be estimated on a per pathline basis; however, it is conceivable that such an approach may fail due to a few reasons: (1) learning a unique $f(\cdot)$ for each pathline can easily result in over-fitting due to the small number of features available, and (2) such an approach fails to take into account any inherent spatial similarity in the error behavior, which can be useful information. Therefore, we train a single model for all pathlines in the flow and exploit a larger set of statistics for a better-fitting error model.

Error prediction using supervised linear regression. The data can be represented as a $p \times (n-1)$ matrix, where errors for $n-1$ successive resolutions are given for p pathlines: each row in the matrix represents errors for a single pathline, and the column j represents the error $\varepsilon_{(j+1,j)}$. The goal is to predict the first column, $\varepsilon_{(2k,k)}$. Since we cannot use the same data for training and validation, and since availing additional data, either in terms of more pathlines or errors at more resolutions, is not possible, we instead train the model on columns $[3, 4, \ldots, n-1]$ to predict column two ($\varepsilon_{(3k,2k)}$). Next, we use the trained model, and predict on columns $[2, 3, \ldots, n-2]$, which we validate against the first column, $\varepsilon_{(2k,k)}$. The underlying hypothesis is that the regression model is able to capture the functional relationship between errors across temporal resolutions, which generalizes well to unseen data. In a realistic scenario, one would train the model using columns $[2, 3, \ldots, n-2]$ and predict the error $\varepsilon_{(k,1)}$

As argued already, we expect the errors for successive sampling to follow an exponential-decay trend. This intuition is supported by our observation that the model gives a better fit in $\log_{10}$ space. Nevertheless, after the first few subsampling steps, (especially the turbulent) pathlines may become significantly erroneous, and a low signal-to-noise ratio may preclude any information from being meaningful. Therefore, we are restricted to a small number of columns (in our experiments, we used $n \leq 10$). In order to improve the (linear) model, we increase the number of features (columns) by including polynomial combinations of existing features, up to degree 2. For example, if $[a, b]$ was the original set of features, we transform them to $[1, a, b, a^2, b^2, ab]$. This standard pre-processing step tends to improve regression performance for machine learning algorithms, as the non-linearity enables the algorithm to approximate more complex relationships similar to the kernel trick [5]. As a result, the size of the feature set becomes 36 (for n=7), upon which we perform training and testing.

Furthermore, since we are changing the columns for testing, we control the difference on the value ranges between the training and test data by normalizing them to $[0, 1]$ through min/max scaling. This scaling controls the variance of the data, and allows the model to be applied to a different set of columns. The dependent variables for training ($\varepsilon_{(3,2)}$) and testing ($\varepsilon_{(2,1)}$) are not scaled.

Finally, we fit a linear regression model to the training data, and predict with the test data. The linear model is chosen for its ease of interpretation, and scalability for large-scale data.

Model evaluation. To evaluate the performance of the model, we use two metrics: (1) the Spearman correlation, which describes how strongly two series of values are correlated, (≈ 1.0 indicates strong correlation), and (2) the R^2 statistic, or the coefficient of determination, which is the proportion of variance in the predicted value that can be explained from the true value, and takes a maximum value of 1.0 to indicate reliable prediction.

4 Validation and Results

Here we use two turbulent flows, one 2D and one 3D, to validate the error estimates discussed above. Using especially-high resolutions, or in the case of the lifted flame, *in situ* pathlines, we demonstrate that our model is able to predict the error due to temporal subsampling reasonably well.

4.1 2D Flow Past a Cylinder

Our first test data is a 2D flow past a cylinder, which was simulated using Nektar++ [8] on a 1300×600 regular grid, with a simulation time-step 0.01 ($\Delta = 1$) and Reynolds number 300. To obtain accurate data for experimentation and validation, snapshots of the flow were saved at an unusually-high frequency: every 10^{th} simulation time-step, *i.e.*, $\Delta = k = 10$. Even for this moderately-sized data, the total size of storing only 600 snapshots of the simulated flow at the chosen resolution amounts to about 3.5 GB, highlighting the challenges in storing finer resolutions. To model subsampling errors, we compute a dense set of pathlines, seeded at every grid point and integrated until they exit the domain. Pathlines with same seed points are computed for subsampled flows, $\vec{V}_{nk}(\mathbf{x}, t)$, for $1 \leq n \leq 10$, and errors between pairs of pathlines are computed at successive resolutions.

Model validation. Since true pathlines ($\Delta = 1$) are not known, we consider $\Delta = k$ as ground truth, and use the model described in the previous section to estimate $\varepsilon_{(2k,k)}$, using data of resolutions $2k$ and coarser only. Fig. 3 shows the density scatter plot of the predicted $\varepsilon_{(2k,k)}$ plotted against the true $\varepsilon_{(2k,k)}$. It indicates a good fit around the ideal 45° line with a slight trend to over-predict for larger errors, as also determined by high values of R2 and Spearman coefficient. Note that the figure uses a logarithmic color map and highlights that the vast majority of pathlines are predicted well and contain errors less than 5 grid cells. Note that this figure provides a zoomed-in view to highlight the details and capture $\approx 99.2\%$ samples.

Error prediction for the given resolution, $\Delta = k$. Finally, we use the model as we would in practice, *i.e.*, to predict the unknown error $\varepsilon_{(k,1)}$, and show the

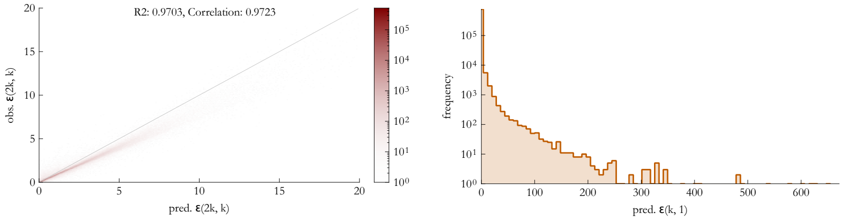

Fig. 3. Left: Validation of our error model for the 2D flow past a cylinder. The model produces good prediction of subsampling errors as shown by the density visualization of correlation between predicted and observed errors. Right: Predicted errors for the 2D flow past a cylinder sampled at the given resolution every 10^{th} time-step. The distribution of predicted error shows that a non-negligible number of pathlines contain large errors, even at this unusually-high sampling.

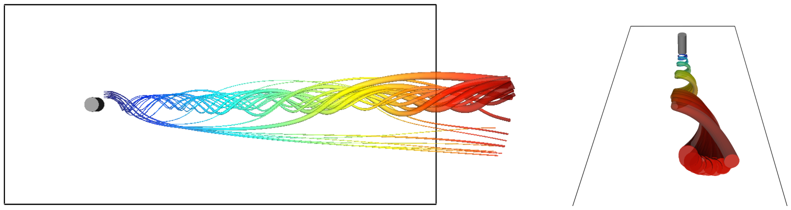

Fig. 4. Uncertainty visualization for the pathlines in the 2D flow past a cylinder. The figures show pathlines as thick tubes by mapping the point-wise error as radii, and time as color. The visualization in (a) shows intersections in these tubes, implying that subsequent FTLE-based analysis may contain arbitrary errors. (b) shows fewer pathlines from a different view point.

resulting histogram in Fig. 3. As seen in the figure, for the vast majority of pathlines, the predicted error is clustered around 0 (note the spike even on a logarithmic scale) as we would expect from the comparatively high temporal resolution. Nevertheless, there exist a relatively-small but not negligible set of pathlines with a predicted error of about 100 grid cells. Even considering the tendency to over-predict, this indicates that despite the high temporal resolution, the stored data contains regions of concern. Imagining, *e.g.*, an FTLE computation with random pathline errors at the scale of multiple grid cells. These types of errors at hundreds of pathlines could create noticeable artifacts.

Uncertainty visualization. We can use these estimations to visualize uncertainty in pathlines. By mapping point-wise errors to radii, we can display pathlines as thick tubes to understand the spatial manifestation of errors in the vicinity of other pathlines. Fig. 4 show such visualizations for selected pathlines, where color represents time and thickness represents estimated error. In particular, the figure shows pathlines seeded from adjacent grid points in a 5×5 neighborhood, which show tubes corresponding to neighboring seed points intersecting. Here, we use the predicted error for a pathline to indicate the final thickness and vary it linearly along the pathline.

4.2 3D Lifted Ethylene Jet Flame

We next study the data presented in Section 1 to evaluate the given sampling rate. As discussed earlier, for the lifted flame, true particle paths are available at a 10× higher frequency than the velocity fields, providing a rare opportunity to validate our technique with highly-accurate simulation data. Nevertheless, since the pathlines errors are already large (see Fig. 2), and the given temporal resolution is already substantially lower, we subsample the data only upto $n = 3$. While this provides much-fewer data for the model, any further subsampling led to substantial artifacts and no longer reasonably approximated the flow.

Comparing the predicted errors for the pathlines at $\Delta = k$, with the computed errors (with respect to the *in situ* particles) leads to the scatter plot in Fig. 5, which shows that the densest parts of the scatterplot lie on the 45° line corroborating that, on an average, our metric estimates the error reasonably well. While the differences between our estimation and the true error has a high variance, as one would expect in such a complex flow, errors in most pathlines are estimated within about 100 grid points at an average error of around 20 grid cells. Even conservatively, one would, therefore, expect a random error in pathlines of about 20 grid cells which would raise significant concerns about the reliability of the underlying pathlines. Fig. 5 shows the residual plot and highlights, once again, that whereas our model relatively over-predicts the error, most pathlines lie near the origin suggesting a good prediction overall.

By mapping the point-wise errors to radii, we show pathlines as thick tubes to visualize how the error evolves along the length of the pathline. As before, the final width of the tubes are as large as the predicted per-pathline error with the width scaled linearly along the length. Fig. 6 shows large errors accumulated near the end of pathlines highlighting the potentially substantial errors in the given resolution and its implication on any subsequent analysis.

5 Conclusion

This paper presents a new *a posteriori* estimate for errors in computation of pathlines due to temporal subsampling of unsteady flows. Whereas the existing error studies for pathline tracing either address other more-amenable sources of errors or require the knowledge of the ground truth and/or the expected time-scales of features in the flow, our technique estimates the error without requiring any prior knowledge about a given flow. Instead, our model directly analyzes relationships between error and temporal resolution for artificially subsampled data to derive error estimates.

Although we do not make any assumptions about the underlying flow and expect this technique to be generally applicable, it is important to better understand how other factors, such as other classes of flow and different integrators, would impact the model. We also assume that the given data is reasonably sampled and, therefore, expect that further subsampling creates a tractable loss of information. In cases where the initial data is already too sparse for meaningful results, further subsampling may not provide useful insights.

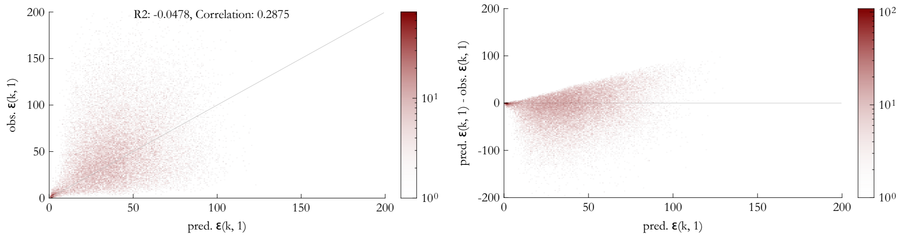

Fig. 5. Prediction of error in the pathlines of lifted flame. The figures correlate the predicted error with the computed error (using insitu-particles) as scatter and residual plots, showing that the prediction has low bias suggesting a good model fit, but contains high variance due to fewer available pathlines and resolutions.

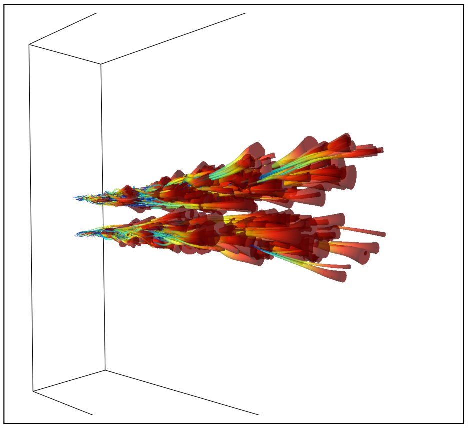

Fig. 6. Visualization of pathlines as tubes with radii mapped to point-wise error describes the evolution of error along the pathline, as well has enables understanding the sources of incorrect analysis in nearby pathlines.

For the lifted flame data, only a small number of pathlines (54,935) are available, *all* of which are turbulent, and the errors are distributed rather randomly (see Fig. 2). As a result, the relationship between the errors at successive resolutions and the relationship between errors of different pathlines are difficult to capture and the model shows high variance, resulting in suboptimal predictions. In comparison, a much-larger set of pathlines is available for training the model for the 2D flow past a cylinder (780,000). Furthermore, many of these pathlines show coherent behavior, *e.g.*, many pathlines are laminar in a similar way, whereas, others produce similar turbulence. Therefore, we see low variance in the prediction. Unsurprisingly, the model produces more-accurate predictions when the training data set is large and coherent, whereas, in other cases, the predictions are less accurate. Moreover, a majority of the pathlines are relatively simple and, therefore, show less error. As a result, the regression model is biased towards seeing such samples. With less training on complex pathlines, the model tends to over-predict. On the other hand, almost all pathlines are turbulent and contain large errors in which case the model provides unbiased results. The problems with creating biased prediction is a known limitations of such a simple model and more advanced regression techniques could likely improve the predictions. However, the results would be less interpretable and more challenging to reproduce. Furthermore, the goal is not necessarily to develop an accurate per-pathline prediction, which, given the chaotic nature of turbulence, is likely an unrealistic goal. Instead, our approach aims to highlight the overall trends to allow a qualitative assessment on which pathlines are likely to reliably represent an underlying flow. The overarching challenge remains in obtaining data that is sampled sufficiently finely, such that, a meaningful model can be constructed through temporal subsampling. Another potential direction of future work could be to reformulate the model as a classification task, where one could predict the error as being one of three classes — low, medium or high. This makes the learning problem more regularized, especially with respect to the extremely-turbulent pathlines. In addition, an added constraint could be to employ a loss function such as the Wasserstein loss [17], or use an ordinal classification framework [16], which enforces the natural ordering of classes (low error < medium error < high error) into the loss.

Discussion and Outlook. The analysis presented above raises serious concerns about the reliability of post-hoc pathlines and their subsequent analysis. Notice that even at impractically high temporal resolutions, the cylinder model suggests that there exist hundreds or even thousands of pathlines with errors beyond five grid cells. Considering that one of the primary reason to compute pathlines is to derive FTLE fields, unstructured errors of this magnitude and beyond are likely to cause severe artifacts. Clearly, the exact impact of such artifacts will depend on the specific uses case, the nature of the flow, as well as a host of other factors. However, this study suggests that evaluating the impact of temporal subsampling should be an integral part of any pathline-based analysis to better understand the inherent uncertainties and potential errors.

Approaches like the one presented here open a number of interesting research directions and provide opportunities to re-engage the broader scientific community with new explicitly validated approaches and reliable error predictions. Furthermore, this work highlights the need to develop better interpolation schemes to reduce the errors or new representations like the Lagrangian one [1] to completely avoid temporal subsampling. In this context, the simple model proposed above represents only a first step in developing more general diagnostics for pathline-based analysis approaches.

Ackwnowledgements. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. This work was funded in part by NSF OAC awards 2127548, 1941085, NSF CMMI awards 1629660, DoE award DE-FE0031880, and the Intel Graphics and Visualization Institute of XeLLENCE. RMK acknowledges support from ARO W911NF-15-1-0222 (Program Manager Dr. Mike Coyle). Release number: LLNL-CONF-826636.

References

1. Agranovsky, A., Camp, D., Garth, C., Bethel, E.W., Joy, K.I., Childs, H.: Improved post hoc flow analysis via Lagrangian representations. In: Proc. of IEEE Symp. Large Data Analysis and Vis. (LDAV). pp. 67–75 (2014)
2. Berger, K., Berger, K., Callet, P.L.: UHD image reconstruction by estimating interpolation error. In: 2015 IEEE International Conference on Image Processing (ICIP). pp. 4743–4747 (Sept 2015)
3. Bhatia, H., Jadhav, S., Bremer, P.T., Chen, G., Levine, J.A., Nonato, L.G., Pascucci, V.: Edge maps: Representing flow with bounded error. In: Proc. of IEEE Pacific Vis. Symp. pp. 75–82 (March 2011)
4. Bhatia, H., Jadhav, S., Bremer, P.T., Chen, G., Levine, J.A., Nonato, L.G., Pascucci, V.: Flow visualization with quantified spatial and temporal errors using edge maps. IEEE Trans. Vis. Comput. Graph. **18**(9), 1383–1396 (Sept 2012)
5. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer-Verlag New York (2006)
6. Braun, S.A., Montgomery, M.T., Pu, Z.: High-resolution simulation of hurricane bonnie (1998). Part I. J. Atmos. Sci. **63**(1), 19–42 (2006)
7. Butcher, J.C.: A history of runge-kutta methods. Appl. Numer. Math. **20**(3), 247–260 (Mar 1996)
8. Cantwell, C., Moxey, D., Comerford, A., Bolis, A., Rocco, G., Mengaldo, G., Grazia, D.D., Yakovlev, S., Lombard, J.E., Ekelschot, D., Jordi, B., Xu, H., Mohamied, Y., Eskilsson, C., Nelson, B., Vos, P., Biotto, C., Kirby, R., Sherwin, S.: Nektar++: An open-source spectral/ element framework. Computer Physics Communications **192**, 205–219 (2015)
9. Cartwright, J.H.E., Feudel, U., Károlyi, G., de Moura, A., Piro, O., Tél, T.: Dynamics of finite-size particles in chaotic fluid flows. In: Nonlinear Dynamics and Chaos: Advances and Perspectives, pp. 51–87. Springer Berlin Heidelberg, Berlin, Heidelberg (Apr 2010)
10. Chandler, J., Obermaier, H., Joy, K.I.: Interpolation-based pathline tracing in particle-based flow visualization. IEEE Trans. Vis. Comput. Graph. **21**(1), 68–80 (Jan 2015)
11. Chen, C.M., Biswas, A., Shen, H.W.: Uncertainty modeling and error reduction for pathline computation in time-varying flow fields. In: Proc. of IEEE Pacific Vis. Symp. pp. 215–222 (April 2015)
12. Chen, J., Choudhary, A., Feldman, S., Hendrickson, B., Johnson, C.R., Mount, R., Sarkar, V., White, V., Williams, D.: Synergistic Challenges in Data-Intensive Science and Exascale Computing: DOE ASCAC Data Subcommittee Report. Department of Energy Office of Science (March 2013), type: Report
13. Chen, J.H., Choudhary, A., de Supinski, B., DeVries, M., Hawkes, E.R., Klasky, S., Liao, W.K., Ma, K.L., Mellor-Crummey, J., Podhorszki, N., Sankaran, R., Shende, S., Yoo, C.S.: Terascale direct numerical simulations of turbulent combustion using S3D. Computational Science & Discovery **2**(1), 015001 (2009)
14. Darmofal, D.L., Haimes, R.: An analysis of 3D particle path integration algorithms. J. Comput. Phys. **123**(1), 182–195 (Jan 1996)
15. Ferstl, F., Bürger, K., Theisel, H., Westermann, R.: Interactive separating streak surfaces. IEEE Trans. Vis. Comput. Graph. **16**(6), 1569–1577 (2010)
16. Frank, E., Hall, M.: A simple approach to ordinal classification. In: European Conference on Machine Learning. pp. 145–156. Springer (2001)

17. Frogner, C., Zhang, C., Mobahi, H., Araya, M., Poggio, T.A.: Learning with a wasserstein loss. In: Advances in Neural Information Processing Systems. pp. 2053–2061 (2015)
18. Günther, T., Kuhn, A., Kutz, B., Theisel, H.: Mass-dependent integral curves in unsteady vector fields. Comput. Graph. Forum **32**(3pt2), 211–220 (Jul 2013)
19. Haller, G.: Finding finite-time invariant manifolds in two-dimensional velocity fields. Chaos **10**(1), 99–108 (2000)
20. Haller, G.: Lagrangian coherent structures and the rate of strain in two-dimensional turbulence. Phys. Fluids A **13**, 3365–3385 (2001)
21. Johnson, C.R., Sanderson, A.R.: A next step: Visualizing errors and uncertainty. IEEE Comp. Graph. and App. **23**(5), 6–10 (Sept 2003)
22. Krishnan, H., Garth, C., Joy, K.I.: Time and streak surfaces for flow visualization in large time-varying data sets. IEEE Trans. Vis. Comput. Graph. **15**(6), 1267–1274 (2009)
23. Maltrud, M., Bryan, F., Peacock, S.: Boundary impulse response functions in a century-long eddying global ocean simulation. Environ. Fluid Mech. **10**, 275–295 (2010)
24. McLoughlin, T., Laramee, R.S., Peikert, R., Post, F.H., Chen, M.: Over two decades of integration-based, geometric flow visualization. Comput. Graph. Forum **29**(6), 1807–1829 (Sep 2010)
25. Otto, M., Germer, T., Theisel, H.: Uncertain topology of 3D Vector Fields. In: Proc. of IEEE Pacific Vis. Symp. pp. 65–74 (2011)
26. Otto, M., Germer, T., Hege, H.C., Theisel, H.: Uncertain 2D vector field topology. Comp. Graph. Forum **29**(2), 347–356 (2010)
27. Pang, A.T., Wittenbrink, C.M., Lodha, S.K.: Approaches to uncertainty visualization. The Visual Computer **13**(8), 370–390 (1996)
28. Peikert, R., Pobitzer, A., Sadlo, F., Schindler, B.: A Comparison of Finite-Time and Finite-Size Lyapunov Exponents. In: Bremer, P.T., Hotz, I., Pascucci, V., Peikert, R. (eds.) Topological Methods in Data Analysis and Visualization III. pp. 187–200. Springer (2014)
29. Teitzel, C., Grosso, R., Ertl, T.: Efficient and reliable integration methods for particle tracing in unsteady flows on discrete meshes. In: Visualization in Scientific Computing. pp. 31–41. Springer (1997)
30. Theisel, H., Weinkauf, T., Hege, H.C., Seidel, H.P.: Topological methods for 2D time-dependent vector fields based on stream lines and path lines. IEEE Trans. Vis. Comput. Graph. **11**(4), 383–394 (July 2005)
31. Üffinger, M., Sadlo, F., Ertl, T.: A time-dependent vector field topology based on streak surfaces. IEEE Trans. Vis. Comput. Graph. **19**(3), 379–392 (2013)
32. Verma, V., Pang, A.T.: Comparative flow visualization. IEEE Trans. Vis. Comput. Graph. **10**(6), 609–624 (2004)
33. Wittenbrink, C.M., Pang, A.T., Lodha, S.K.: Glyphs for visualizing uncertainty in vector fields. IEEE Trans. Vis. Comput. Graph. **2**(3), 266–279 (1996)
34. Yoo, C.S., Richardson, E.S., Sankaran, R., Chen, J.H.: A DNS study on the stabilization mechanism of a turbulent lifted ethylene jet flame in highly-heated coflow. Proc. Combust. Inst. **33**(1), 1619–1627 (2011)
35. Yoo, C.S., Sankaran, R., Chen, J.H.: Three-dimensional direct numerical simulation of a turbulent lifted hydrogen jet flame in heated coflow: Flame stabilization and structure. J. Fluid Mech. **640**, 453–481 (2009)