Leveraging information storage to select forecast-optimal parameters for delay-coordinate reconstructions

Joshua Garland, Ryan G. James, and Elizabeth Bradley

Phys. Rev. E **93**, 022221 — Published 29 February 2016

DOI: 10.1103/PhysRevE.93.022221

A new method for choosing parameters in delay reconstruction-based forecast strategies

Joshua Garland*
University of Colorado
Department of Computer Science
Boulder, Colorado 80303, USA

Ryan G. James†
University of California
Department of Physics, Davis, California 95616, USA

Elizabeth Bradley‡
University of Colorado
Department of Computer Science,
Boulder, Colorado 80303, USA
and the Santa Fe Institute, Santa Fe, New Mexico

ABSTRACT

Delay-coordinate reconstruction is a proven modeling strategy for building effective forecasts of nonlinear time series. The first step in this process is the estimation of good values for two parameters, the time delay and the embedding dimension. Many heuristics and strategies have been proposed in the literature for estimating these values. Few, if any, of these methods were developed with forecasting in mind, however, and their results are not optimal for that purpose. Even so, these heuristics—intended for other applications—are routinely used when building delay coordinate reconstruction-based forecast models. In this paper, we propose a new strategy for choosing optimal parameter values for forecast methods that are based on delay-coordinate reconstructions. The basic calculation involves maximizing the shared information between each delay vector and the future state of the system. We illustrate the effectiveness of this method on several synthetic and experimental systems, showing that this metric can be calculated quickly and reliably from a relatively short time series, and that it provides a direct indication of how well a near-neighbor based forecasting method will work on a given delay reconstruction of that time series. This allows a practitioner to choose reconstruction parameters that avoid any pathologies, regardless of the underlying mechanism, and maximize the predictive information contained in the reconstruction.

I. INTRODUCTION

The method of delays is a well-established technique for reconstructing the state-space dynamics of a system from scalar time-series data[1–3]. The task of choosing good values for the free parameters in this procedure has been the subject of a large and active body of literature over the past few decades, e.g.,[4–15]. The majority of these techniques focus on the geometry of the reconstruction. A standard method for selecting the delay τ, for instance, is to maximize independence between the coordinates of the delay vector while minimizing overfolding and reduction in causality between coordinates[5]; a common way to choose an embedding dimension is to track changes in near-neighbor relationships in reconstructions of different dimensions[14].

This heavy focus on the geometry of the delay reconstruction is appropriate when one is interested in quantities like fractal dimension and Lyapunov exponents, but it is not necessarily the best approach when one is building a delay reconstruction *for the purposes of prediction*. That issue, which is the focus of this paper, has received comparatively little attention in the extensive literature on delay reconstruction-based prediction[16–21]. In the following section, we propose a robust, computationally efficient method called *time delayed active information storage*, $\mathcal{A}_\tau$, that can be used to select parameter values that maximize the shared information between the past and the future—or, equivalently, that maximize the reduction in uncertainty about the future given the current model of the past. The implementation details, and a complexity analysis of the algorithm, are covered in Section III. In Section IV, we show that simple prediction methods working with $\mathcal{A}_\tau$-optimal reconstructions—i.e., constructions using parameter values that follow from the $\mathcal{A}_\tau$ calculations—perform better, on both real and synthetic examples, than those same forecast methods working with reconstructions that are built using the traditional methods mentioned above. Finally, in Section V we explore the utility of $\mathcal{A}_\tau$ in the face of different data lengths and prediction horizons.

* joshua.garland@colorado.edu
† rgjames@ucdavis.edu
‡ lizb@colorado.edu

II. SHARED INFORMATION AND DELAY RECONSTRUCTIONS

The information shared between the past and the future is known as the excess entropy[22]. We will denote it here by $E = I[\overleftarrow{X}; \overrightarrow{X}]$, where I is the mutual information[23] and $\overleftarrow{X}$ and $\overrightarrow{X}$ represent the infinite past and the infinite future, respectively. E is often difficult to estimate from data due to the need to calculate statistics over potentially infinite random variables[24]. While this is possible in principle, it is too difficult in practice for all but the simplest of dynamics[25]. In any case, the excess entropy is not exactly what one needs for the purposes of prediction, since it is not realistic to expect to have the infinite past or to predict infinitely far into the future. For our purposes, it is more productive to consider the information contained in the *recent* past and determine how much that explains about the not-too-distant future. To that end, we define the *state active information storage*:

$$\mathcal{A}_{\mathcal{S}} = I[\mathcal{S}_j; X_{j+p}] \ ,$$

where $\mathcal{S}_j$ is an estimate of the state of the system at time j and X_{j+p} is the state of the system p steps in the future. In the case that the state estimate $\mathcal{S}$ takes the form of a delay vector with delay τ, we will refer to it as the *time delayed active information storage* and use the symbol $\mathcal{A}_\tau$.

This can be neatly visualized—and compared to traditional methods like time-delayed mutual information, multi-information and the so-called co-information[26]—using the I-diagrams of Yeung[23]. Figure 1 shows an I-diagram of time-delayed mutual information for a specific τ. In a diagram like this, each circle represents the uncertainty in a particular variable. The left circle in Figure 1, for instance, represents the average uncertainty in observing $X_{j-\tau}$ (i.e., $H[X_{j-\tau}]$, where H is the Shannon entropy[23]); similarly, the top circle represents $H[X_{j+p}]$, the uncertainty in the p^{th} future observation. Each of the overlapping regions represents *shared* uncertainty: e.g., in Figure 1, the shaded region represents the shared uncertainty between X_j and $X_{j-\tau}$. More precisely, the shaded region schematizes the quantity

$$\begin{aligned} I[X_j; X_{j-\tau}] &= H[X_j] + H[X_{j-\tau}] - H[X_j, X_{j-\tau}] \\ &= H[X_j] - H[X_j|X_{j-\tau}] \\ &= H[X_{j-\tau}] - H[X_{j-\tau}|X_j]. \end{aligned}$$

If the X are trajectories in reconstructed state space, then tuning the reconstruction parameters (e.g., τ) changes the size of the overlap regions—i.e., the amount of information shared between the coordinates of the delay vector. This notion can be put into practice to select good values for those parameters. Notice, for instance, that minimizing the shaded region in Figure 1—that is, rendering X_j and $X_{j-\tau}$ as independent as possible—maximizes the total uncertainty that is explained by the combined model $[X_j, X_{j-\tau}]^T$ (the sum of the area of the two circles). This is precisely the argument made by Fraser and Swinney in [5]. However, it is easy to see from the I-diagram that choosing τ in this way does not explicitly take into account explanations of the *future*—that is, it does not reduce the uncertainty about X_{t+p}. Moreover, the calculation does not extend to three or more variables, where minimizing overlap is not a trivial extension of the reasoning captured in the I-diagrams.

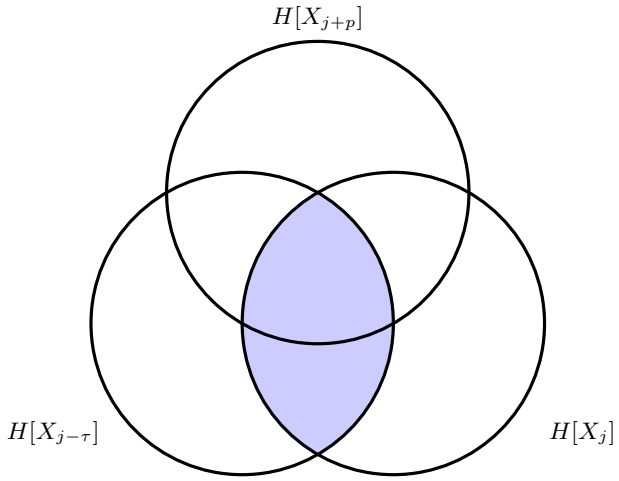


FIG. 1: An I-diagram of the time-delayed mutual information. The circles represent uncertainties (H) in different variables; the shaded region represents $I[X_j; X_{j-\tau}]$, the time-delayed mutual information between the current state X_j and the state τ time units in the past, $X_{j-\tau}$. Notice that the shaded region is indifferent to $H[X_{j+p}]$, the uncertainty about the future.

The obvious next step would be to explicitly include the future in the estimation procedure. One approach to this would be to work with the so-called co-information[26],

$$\mathcal{C} = I[X_j; X_{j-\tau}; X_{j+p}] \ ,$$

As depicted in Figure 2a, this is the intersection of $H[X_j]$, $H[X_{j-\tau}]$ and $H[X_{j+p}]$. It describes the reduction in uncertainty that the *two* past states, together, provide regarding the future. While this is obviously an improvement over the time-delayed mutual information of Figure 1, it does not take into account the information that is shared between X_j and the future but *not shared with the past* (i.e., $X_{j-\tau}$), and vice versa. The so-called multi-information,

$$\mathcal{M} = \sum_{i \in \{j, j-\tau, j+p\}} (H[X_i]) - H[X_j, X_{j-\tau}, X_{j+p}] \ , \qquad (1)$$

depicted in Figure 2b addresses this shortcoming, but it also includes information that is shared between the past and the present, but not with the future. This is

not terribly useful for the purposes of prediction. Moreover, the multi-information overweights information that is shared between all three circles—past, present, and future—thereby artificially over-valuing information that is shared in all delay coordinates. In the context of predicting X_{t+p}, the provenance of the information is irrelevant and so the multi-information also seems ill-suited to the task at hand.

More generally, the multi-information has been used in a similar manner to the time-delayed mutual information above in estimating τ: e.g., for a three dimensional embedding, attempting to minimize $\mathcal{M}[X_j; X_{j-\tau}; X_{j-2\tau}]$. As can be seen in Eq. 1, minimizing the multi-information is equivalent to maximizing the entropy, and with a maximal entropy, the delay vectors are in some sense maximally informative because dependencies among the dimensions have been minimized. While on the surface this may seem a boon to prediction, consider the issue of predicting the state of the system at time $j + \tau$: if the coordinates of the delay vector are maximally independent, they will also be independent of the value being predicted. In light of this, we can conclude that the minimal multi-information approach is not well aligned with the goal of prediction.

$\mathcal{A}_\tau$ addresses all of the issues raised in the previous paragraphs. By treating the generic delay vector as a joint variable, rather than a series of single variables, $\mathcal{A}_\tau$ captures the shared information between the past, present, and future independently (the left and right colored wedges in Figure 3), as well as the information that the past and present, together, share with the future (the center wedge). By choosing delay reconstruction parameters that maximize $\mathcal{A}_\tau$, then, one can explicitly maximize the amount of information that each delay vector contains about the future.

To make all of this more concrete and tie it back to state-space prediction of dynamical systems, consider the following example: let $\mathcal{S}_j$ be a two-dimensional delay reconstruction of the time series, $\mathcal{S}_j = [x_j, x_{j-\tau}]^T$. In this case, $\mathcal{A}_\tau$ becomes $I[[X_j, X_{j-\tau}]^T; X_{t+p}]$, which describes the reduction in uncertainty about the system at time $j + p$, given the state estimate $[X_j, X_{j-\tau}]^T$. One can estimate a τ value for the purposes of reconstructing the dynamics from a given time series, for instance, by calculating $\mathcal{A}_\tau$ for a range of τ and choosing the first maximum (i.e., minimizing the uncertainty about the p^{th} future observation). One can then apply any state-space forecasting method to the resulting reconstruction in order to predict the future course of that time series. In Section IV, we explore that claim using Lorenz's classic method of analogues[21], but it should be just as applicable for other predictors that utilize state-space reconstructions, such as the methods used in[16–18, 20].

Notice that both the definition of the state active information storage, $\mathcal{A}_\mathcal{S}$, and its use in optimizing forecast algorithms are general ideas that are easily extensible to other state estimators. For example, in the case of traditional delay-coordinate *embedding*, the state estimator

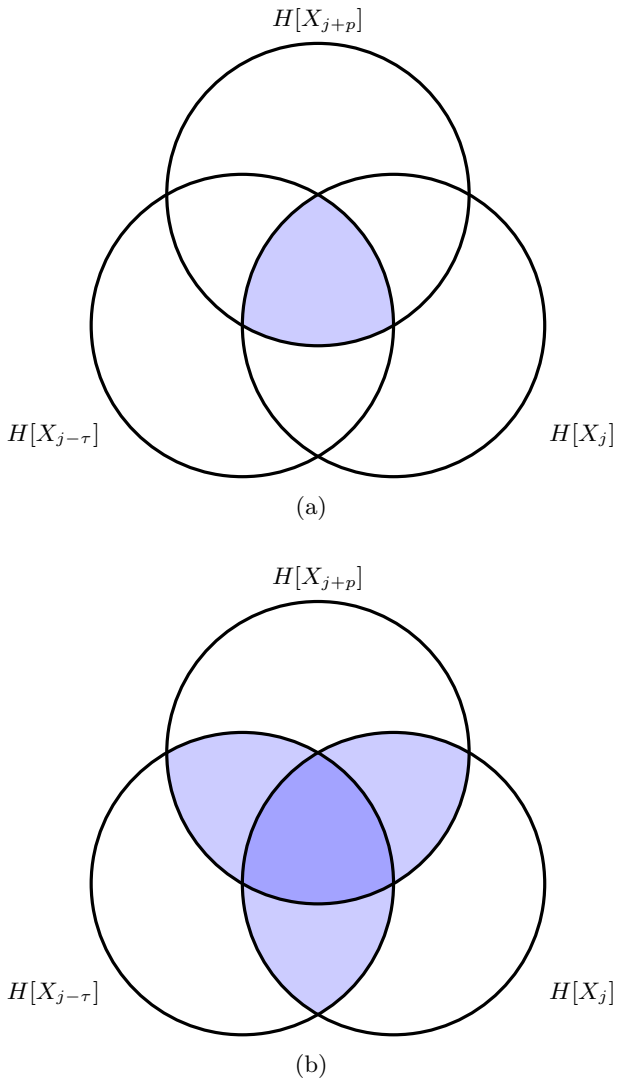


FIG. 2: Two possible generalizations of the mutual information. (a) The co-information, $\mathcal{C}[X_{j+1}; X_j; X_{j-\tau}]$. (b) An I-diagram of the multi-information, $\mathcal{M}[X_j, X_{j-\tau}; X_{j+p}]$. The centermost region is more darkly shaded here to reflect the extra weight that that region carries in the calculation.

is the m-dimensional delay vector, i.e.,

$$\mathcal{S}_j = [X_j, X_{j-\tau}, \ldots, X_{j-(m-1)\tau)}]^T$$

with m chosen to meet the appropriate theoretical requirements [1, 3], resulting in our $\mathcal{A}_\tau$. We demonstrate this approach in Section IV. If the time series is preprocessed (e.g., via a Kalman filter[27], a low-pass filter and an inverse Fourier transform[28], or some other local linear transformation[6, 16–18, 20]), the state estimator simply becomes $\mathcal{S}_j = \hat{\vec{x}}_j$ where $\hat{\vec{x}}_j$ is the processed m-dimensional delay vector. As we demonstrate in Section IV B, one can even use $\mathcal{A}_\mathcal{S}$ to optimize parameter choices for forecast methods that use reconstructions that are not embeddings—i.e., those whose dimensions do not

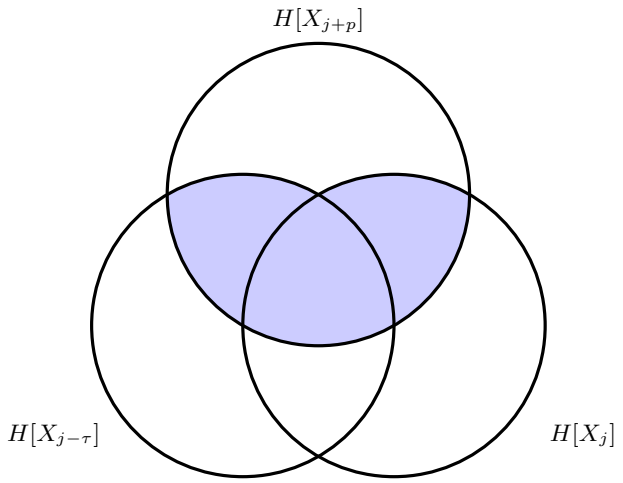


FIG. 3: An I-diagram of $\mathcal{A}_\tau$, the quantity proposed in this paper: $I[[X_j, X_{j-\tau}]; X_{j+p}]$. This quantity captures the shared information between the past, present, and future independently, as well as the information that the past and present, together, share with the future.

meet the traditional requirements for preserving dynamical invariants like the Lyapunov exponent.

III. EFFICIENT ESTIMATION OF $\mathcal{A}_\tau$

To calculate $\mathcal{A}_\tau$ from a real-valued time series, one must first symbolize those data. Simple binning is not a good solution here, as it is known to cause severe bias if the bin boundaries do not create a generating partition[29]. A useful alternative is kernel estimation [30, 31], in which the relevant probability density functions are estimated via a function Θ with a resolution or bandwidth r that measures the similarity between two points in $X \times Y$ space. (For $\mathcal{A}_\tau$, X would be $\mathcal{S}_j$ and Y would be X_{j+p}.) Given points $\{x_i, y_i\}$ and $\{x'_i, y'_i\}$ in $X \times Y$, one can define:

$$\hat{p}_r(x_i, y_i) = \frac{1}{N} \sum_{i'=1}^{N} \Theta\left(\left| \begin{matrix} x_i - x'_i \\ y_i - y'_i \end{matrix} \right| - r \right) ,$$

where $\Theta(x > 0) = 0$ and $\Theta(x \leq 0) = 1$. That is, $\hat{p}_r(x_i, y_i)$ is the proportion of the N pairs of points in $X \times Y$ space that fall within the kernel bandwidth r of $\{x_i, y_i\}$, i.e., the proportion of points similar to $\{x_i, y_i\}$. When $|\cdot|$ is the max norm, this is the so-called box kernel. This too, however, can introduce bias[32] and is dependent on the choice of bandwidth r. After these estimates, and the analogous estimates for $\hat{p}(x)$, are produced, they are then used directly to compute local estimates of mutual information for each point in space, which are then averaged over all samples to produce the mutual information of the time series. For more details on this procedure, see[32].

A better way to calculate $I[X;Y]$ and estimate $\mathcal{A}_\tau$ is the Kraskov-Stügbauer-Grassberger (KSG) estimator[29]. This approach dynamically alters the kernel bandwidth to match the density of the data, thereby smoothing out errors in the probability density function estimation process. In this approach, one first finds the k^{th} nearest neighbor for each sample $\{x, y\}$ (using max norms to compute distances in x and y), then sets kernel widths r_x and r_y accordingly and performs the pdf estimation. There are two algorithms for computing $I[X;Y]$ with the KSG estimator[32]. The first is more accurate for small sample sizes but more biased; the second is more accurate for larger sample sizes. We use the second of the two in the results reported in this paper, as we have fairly long time series. Our algorithm sets r_x and r_y to the x and y distances to the k^{th} nearest neighbor. One then counts the number of neighbors within and on the boundaries of these kernels in each marginal space, calling these sums n_x and n_y, and finally calculates

$$I[X;Y] = \psi(k) - \frac{1}{k} - \langle \psi(n_x) + \psi(n_y) \rangle + \psi(n) ,$$

where ψ is the digamma function[33]. This estimator has been demonstrated to be robust to variations in k as long as $k \geq 4$[32].

In this paper, we employ the Java Information Dynamics Toolkit (JIDT) implementation of the KSG estimator[32]. The computational complexity of this implementation is $\mathcal{O}(kN \log N)$, where N is the length of the time series and k is the number of neighbors being used in the estimate. While this is more expensive than traditional binning ($\mathcal{O}(N)$), it is bias corrected, allows for adaptive kernel bandwidth to adjust for under- and over-sampled regions of space, and is both model and parameter free (aside from k, to which it is very robust).

IV. APPLYING $\mathcal{A}_\tau$ TO SELECT RECONSTRUCTION PARAMETERS

In this section, we demonstrate how to use $\mathcal{A}_\tau$ to choose parameter values for delay-reconstruction forecast models. We do this for several synthetic examples, as well as for sensor data from several laboratory experiments. For the discussion that follows, we use the term "$\mathcal{A}_\tau$-optimal" to refer to the parameter values (m and τ) that provided the best match between the forecast and the true continuation.

To evaluate a forecast model, we divide the signal into two parts: the initial training signal $\{x_j\}_{j=1}^{n}$—the first n elements of the time series—and the test signal $\{c_\ell\}_{\ell=n+1}^{k+n+1}$, where k is the length of the prediction. We build a delay reconstruction from the x_j (i.e., a sequence of points $[x_j, x_{j-\tau}, \ldots, x_{j-(m-1)\tau)}]^T$), use it to generate a prediction $\{\hat{x}_\ell\}_{\ell=n+1}^{k+n+1}$, and then use the Mean Absolute Scaled Error[34] to compare the prediction to the test signal:

$$MASE = \sum_{\ell=n+1}^{k+n+1} \frac{|\hat{x}_\ell - c_\ell|}{\frac{k}{n-1}\sum_{j=2}^{n} |x_j - x_{j-1}|}$$

$MASE$ is a normalized measure: the scaling term in the denominator is the average in-sample forecast error for a random-walk prediction—which uses the previous value in the observed signal as the forecast—calculated over the training signal. That is, $MASE < 1$ means that the prediction error in question was, on the average, smaller than the in-sample error of a random-walk forecast on the training portion of the same data. Analogously, $MASE > 1$ means that the corresponding prediction method did *worse*, on average, than the random-walk method.

While its comparative nature may seem odd, this error metric allows for fair comparison across varying methods, prediction horizons, and signal scales, making it a standard error measure in the forecasting literature—and a good choice for the study described in the following sections, which involve a number of very different signals.

A. Synthetic examples

In this Section, we apply $\mathcal{A}_\tau$ to some standard synthetic examples, both maps (Hénon, logistic) and flows: the classic Lorenz system[35] and the more-recent "Lorenz 96" atmospheric model[36]. We construct the traces for the Lorenz experiments using a standard fourth-order Runge-Kutta solver on the associated differential equations, with a timestep of $\frac{1}{64}$, for 60,000 time steps. For the maps, we simply iterate the difference equations 60,000 times. In all cases, we discard the first 10,000 points of each trajectory to remove transient behavior, then sample individual state variables to produce different scalar time-series data sets. We reconstruct the dynamics from those traces using different values of the dimension m and delay τ and compute $\mathcal{A}_\tau$ for each of those reconstructed trajectories. We then use Lorenz's classic method of analogues (LMA) [21] to generate forecasts of each trace, compute their $MASE$ scores as described above, and discuss their relationships to the $\mathcal{A}_\tau$ values for the corresponding time series. For simplicity, in this initial discussion we perform a series of one-step-ahead predictions, rebuilding the model at each step. For the $\mathcal{A}_\tau$ calculations, this means that we estimate $I[\mathcal{S}_j, X_{j+1}]$, with $\mathcal{S}_j = [X_j, X_{j-\tau}, \ldots, X_{j-(m-1)\tau)}]^T$. In Section V B we expand this discussion by increasing the prediction horizon; in Section V A, we consider the effects of the length of the traces.

Flow examples

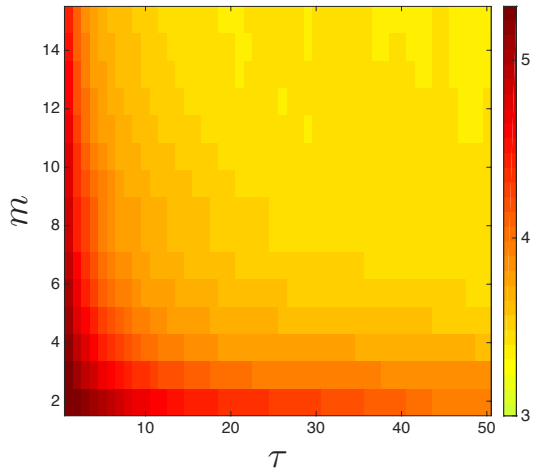
The Lorenz 96 system[36] is defined by a set of K differential equations in the state variables $\xi_1 \ldots \xi_K$:

$$\dot{\xi}_k = (\xi_{k+1} - \xi_{k-2})(\xi_{k-1}) - \xi_k + F$$

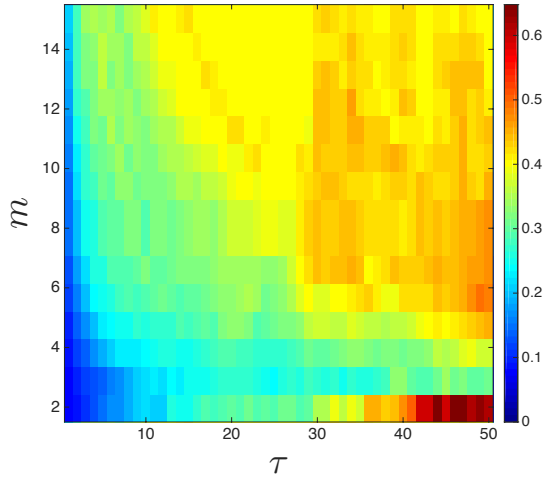
for $k = 1, \ldots, K$, where $F \in \mathbb{R}$ is a constant forcing term that is independent of k. In the following discussion we focus on two parameter sets, $\{K = 22, F = 5\}$ and $\{K = 47, F = 5\}$, which produce low- and high-dimensional chaos, respectively. See [37] for an explanation of this model and the associated parameters.

Figure 4a shows a heatmap of the $\mathcal{A}_\tau$ values for reconstructions of a representative trajectory from this system with $\{K = 22, F = 5\}$, for a range of m and τ. Not surprisingly, this image reveals a strong dependency between the values of the reconstruction parameters and the reduction in uncertainty about the near future that is provided by the reconstruction. Very low τ values, for instance, produce delay vectors that have highly redundant coordinates, and so provide substantial information about the immediate future. As mentioned in the first Section of this paper, standard heuristics only focus on minimizing redundancy between coordinates, choosing the τ value that minimizes the mutual information between the first two coordinates in the delay vector. For this Lorenz 96 trajectory, the approach of Fraser & Swinney [5] yields $\tau = 26$, while standard dimension-estimation heuristics [14] suggest $m = 8$. The $\mathcal{A}_\tau$ value for a delay reconstruction built with those parameter values is 3.463. This is *not*, however, the $\mathcal{A}_\tau$-optimal reconstruction; choosing $m = 2$ and $\tau = 1$, for instance, results in a higher value ($\mathcal{A}_\tau = 5.303$)—i.e., signficantly more reduction in uncertainty about the future. This may be somewhat counter-intuitive, since each of the delay vectors in the $\mathcal{A}_\tau$-optimal reconstruction spans far less of the data set and thus one would expect points in that space to contain *less* information about the future. Figure 4a suggests, however, that this in fact not the case; rather, that uncertainty *increases* with both dimension and time delay.

The question at issue in this paper is whether that reduction in uncertainty about the future correlates with improved accuracy of an LMA forecast built from that reconstruction. Since the $\mathcal{A}_\tau$-optimal choices maximize the shared information between the state estimator and X_{j+1}, one would expect a delay reconstruction model built with those choices to afford LMA the most leverage. To test that conjecture, we performed an exhaustive search with $m = 2, \ldots, 15$ and $\tau = 1, \ldots, 50$. For each $\{m, \tau\}$ pair, we used LMA to generate forecasts from the corresponding reconstruction, computed their $MASE$ scores, and plotted the results in a heatmap similar to the one in Figure 4a. As one would expect, the $MASE$ and $\mathcal{A}_\tau$ heatmaps are generally antisymmetric. This antisymmetry breaks down somewhat for low m and high τ, where the forecast accuracy is low even though



(a)



(b)

FIG. 4: The effects of reconstruction parameter values on $\mathcal{A}_\tau$ and forecast accuracy for the Lorenz 96 system. (a) $\mathcal{A}_\tau$ values for different delay reconstructions of a representative trace from the Lorenz 96 system with $\{K = 22, F = 5\}$. (b) MASE scores for LMA forecasts on different delay reconstructions of a representative trace of the Lorenz 96 system with $\{K = 22, F = 5\}$.

the reconstruction contains a lot of information about the future.

We suspect that this is due to a combination of overfolding (due to too-large values of τ) and projection (low m). Even though each point in such a reconstuction may contain a lot of information about the future, the false crossings created by this combination of effects pose problems for a near-neighbor forecast strategy like LMA. The improvement that occurs if one adds another dimension is consistent with this explanation. Notice, too, that this effect only occurs far from the maximum in the $\mathcal{A}_\tau$ surface—the area that is of interest if one is using $\mathcal{A}_\tau$ to choose parameter values for reconstruction models.

In general, though, maximizing the redundancy between the state estimator and the future does appear to minimize the resulting forecast error of LMA. Indeed, the maximum on the surface of Figure 4b ($m = 2, \tau = 1$) is exactly the minimum on the surface of Figure 4a. The accuracy of this forecast is more than five times higher ($MASE = 0.0737$) than that of a forecast constructed with the parameter values suggested by the standard heuristics (0.3787). Note that the optima of these surfaces may be broad: i.e., there may be *ranges* of m and τ for which $\mathcal{A}_\tau$ and $MASE$ are optimal, and roughly constant. In these cases, it makes sense to choose the lowest m on the plateau, since that minimizes computational effort, data requirements, and noise effects; see [38] for a full discussion of this.

While the results discussed in the previous paragraph do provide a preliminary validation of the claim that one can use $\mathcal{A}_\tau$ to select good parameter values for delay reconstruction-based forecast strategies, they only involve a single example system. Similar experiments on traces from the Lorenz 96 system with different parameter values $\{K = 47, F = 5\}$ (not shown) demonstrate identical results—indeed, the heatmaps are visually indistinguishable from the ones in Figure 4. Figure 5 shows heatmaps of $\mathcal{A}_\tau$ and $MASE$ for similar experiments on the classic "Lorenz 63" system[35]:

$$\begin{aligned}\dot{x} &= \sigma(y - x)\\ \dot{y} &= x(\rho - z) - y\\ \dot{z} &= xy - \beta z\end{aligned}$$

with the typical chaotic parameter selections: $\rho = 28, \sigma = 10$, and $\beta = 8/3$. As in the Lorenz 96 case, the heatmaps are generally antisymmetric, confirming that maximizing $\mathcal{A}_\tau$ is roughly equivalent to minimizing $MASE$. Again, though, the antisymmetry is not perfect; for high τ and low m, the effects of projecting an overfolded attractor cause false crossings that trip up LMA. As before, adding a dimension mitigates this effect by removing these false crossings. Both the Lorenz 63 and Lorenz 96 plots show a general decrease in predictability for large m and high τ, with roughly hyperbolic equipotentials dividing the colored regions[39]. The locations and heights of these equipotentials differs because the two signals are not equally easy to predict. This matter is discussed further at the end of this section.

Numerical $\mathcal{A}_\tau$ and $MASE$ values for LMA forecasts on different reconstructions of both Lorenz systems are tabulated in the top three rows of Table I, along with the reconstruction parameter values that produced those results. The data in this table bring out two important points. First, as suggested by the heatmaps, the m and τ values that maximize $\mathcal{A}_\tau$ (termed $m_{\mathcal{A}_\tau}$ and $\tau_{\mathcal{A}_\tau}$ in the table legend) are close, or identical, to the values that minimize $MASE$ (m_E and τ_E) for all three Lorenz systems. This is notable because—as discussed in Section V A—the former can be estimated quite reliably from a small

TABLE I: $MASE$ values for various delay reconstructions of the different examples studied here. $MASE_H$ is the representative accuracy of LMA forecasts that use delay reconstructions with parameter values ($m_{\mathcal{A}_\tau}$ and $\tau_{\mathcal{A}_\tau}$) chosen via standard heuristics for the corresponding traces—the methods of false neighbors [14] and time-delayed mutual information [5], respectively. Similarly, $MASE_{\mathcal{A}_\tau}$ is the accuracy of LMA forecasts that use reconstructions built with the m and τ values that maximize $\mathcal{A}_\tau$, and $MASE_E$ is the error of the best forecasts for each case, found via exhaustive search over the m, τ parameter space. **: on these signals the standard heuristics failed.

Signal	$MASE_H$	τ_H	m_H	$MASE_{\mathcal{A}_\tau}$	$\tau_{\mathcal{A}_\tau}$	$m_{\mathcal{A}_\tau}$	$MASE_E$	τ_E	m_E
Lorenz-96 $K = 22$	0.3787	26	8	0.0737	1	2	0.0737	1	2
Lorenz-96 $K = 47$	1.007	31	10	0.1156	1	2	0.1156	1	2
Lorenz 63	0.2215	12	5	0.0509	1	3	0.0506	1	2
Hénon Map	**	**	**	3.814e-04	1	2	3.814e-04	1	2
Logistic Map	**	**	**	1.680e-05	1	1	1.680e-05	1	1

sample of the trajectory in only a few seconds of compute time, whereas the exhaustive search that is involved in computing m_E and τ_E for Table I required close to 30 hours of CPU time per signal. A second important point that is apparent from the Table is that delay reconstructions built using the traditional heuristics—the values with the H subscript—were comparatively ineffective for the purposes of LMA-based forecasting. This is notable because that is the default approach in the literature on state-space based forecasting methods for dynamical systems.

A close comparison of Figures 4 and 5 brings up another important point: some time series are harder to forecast than others. Figure 6 breaks down the details of the two suites of Lorenz-96 experiments, showing the distribution of $\mathcal{A}_\tau$ and $MASE$ values for all of the reconstructions. Although there is some overlap in the $K = 47$ and $K = 22$ histograms—i.e., best-case forecasts of the former are better than most of the forecasts of the latter—the $K = 47$ traces generally contain less information about the future and thus are harder to forecast accurately.

Map examples

Delay reconstruction of discrete-time dynamical systems, while possible in theory, can be problematic in practice. Although the embedding theorems do apply in these cases, the heuristics for estimating m and τ often fail. The time-delayed mutual information of [5], for example, may decay exponentially, without showing any clear minimum. And the lack of spatial continuity of the orbit of a map violates the underlying idea behind the method of [14]. State space-based forecasting methods can, however, be very useful in generating predictions of trajectories from systems like this—*if* one has a reconstruction that is faithful to the true dynamics.

In view of this, it would be particularly useful if one could use $\mathcal{A}_\tau$ to choose embedding parameter values for maps. This section explores that notion using two canonical examples, shown in the bottom two rows of Table I. For the Hénon map,

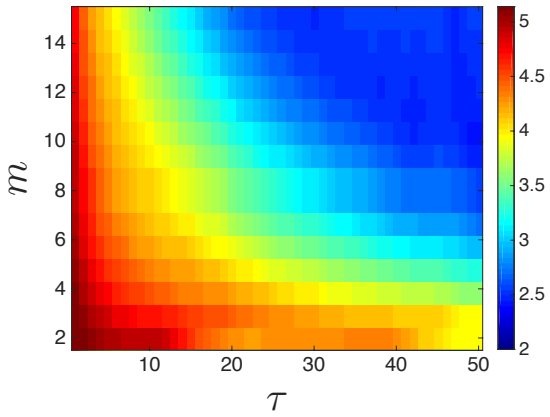
$$x_{n+1} = 1 - ax_n^2 + y_n$$
$$y_{n+1} = bx_n$$

with $a = 1.4$ and $b = 0.3$, the $\mathcal{A}_\tau$-optimal parameter values were $m = 2$ and $\tau = 1$. As in the flow examples, these were identical to the values that minimized $MASE$. These parameter values make sense, of course; a first-return map of the x coordinate is effectively the Hénon map, so $[x_j, x_{j-1}]$ is a perfect state estimator (up to a scaling term). But in practice, of course, one rarely knows the underlying dynamics of the system that generated a time series, so the fact that one can choose good reconstruction parameter values by maximizing $\mathcal{A}_\tau$ is notable—especially since standard heuristics for that purpose fail in this system.

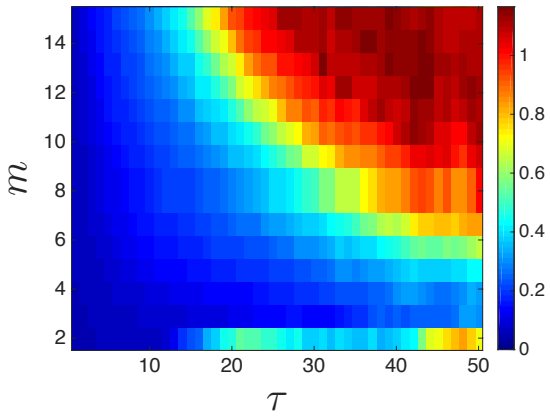
The same pattern holds for the logistic map, $x_{n+1} = rx_n(1 - x_n)$, with $r = 3.65$: the $\mathcal{A}_\tau$-optimal parameter values coincide with the minimum of the $MASE$ surface. As in the Hénon example, these values ($m = 1$ and $\tau = 1$) make complete sense, given the form of the map. But again, one does not always know the form of the system that generated a given time series. In the case of the logistic map, the standard heuristics fail, but $\mathcal{A}_\tau$ clearly indicates that one does not actually need to reconstruct these dynamics—rather, near-neighbor forecasting *on the time series itself* is the best approach.

B. Selecting reconstruction parameters of experimental time series

The results in the previous section provide a preliminary verification of the conjecture that maximizing $\mathcal{A}_\tau$ minimizes forecast accuracy of LMA, for both maps and flows. While experiments with synthetic examples are useful, they do not call the really important aspect of that research question: whether $\mathcal{A}_\tau$ is a useful way to

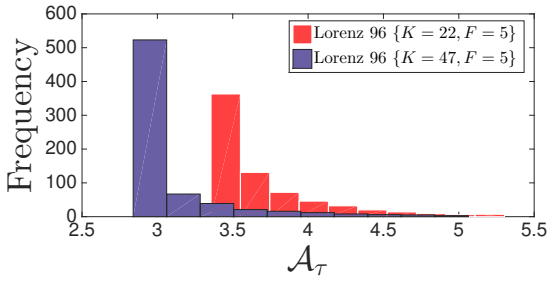


(a)

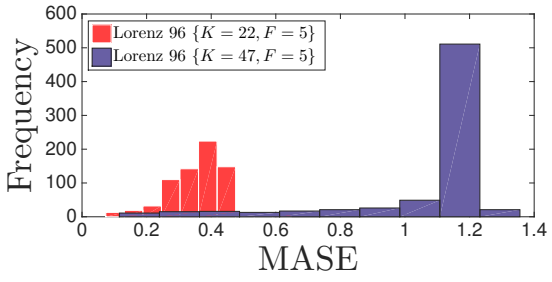


(b)

FIG. 5: The effects of reconstruction parameter values on $\mathcal{A}_\tau$ and forecast accuracy for the Lorenz 63 system. (a) $\mathcal{A}_\tau$ values for different delay reconstructions of a representative trace from the Lorenz 63 system. (b) $MASE$ scores for LMA forecasts on different delay reconstructions of a representative trace of the Lorenz 63 system.



(a)



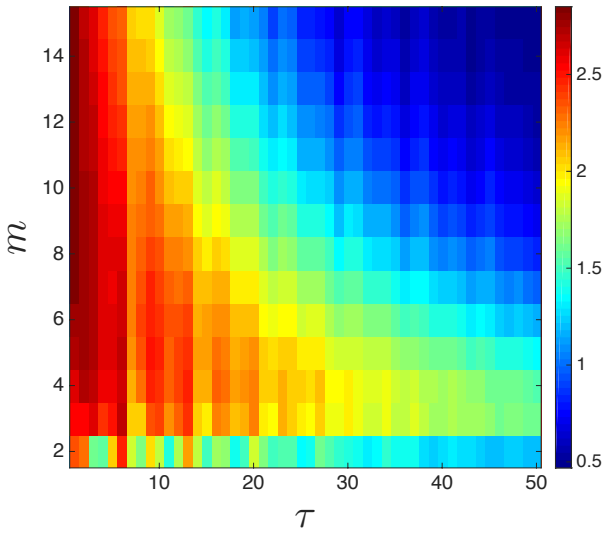
(b)

FIG. 6: Histograms of $\mathcal{A}_\tau$ and $MASE$ values for representative traces from the Lorenz 96 $\{K = 22, F = 5\}$ and $\{K = 47, F = 5\}$ systems for all $\{m, \tau\}$ values in Figures 4 and 5. (a) $\mathcal{A}_\tau$. (b) $MASE$.

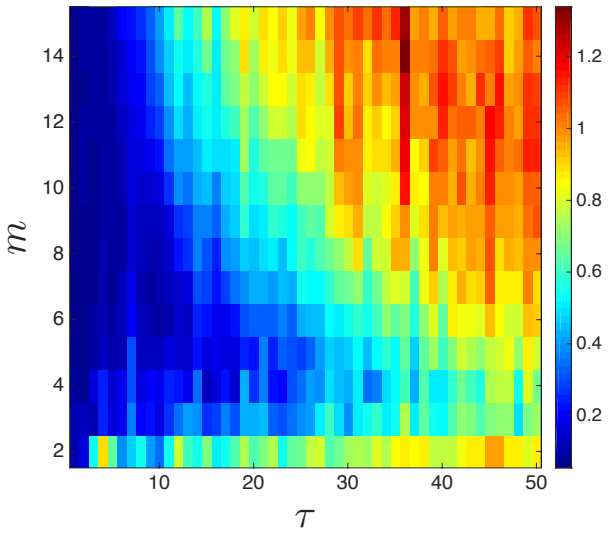
choose parameter values for delay reconstruction-based forecasting of real-world data, where the time series are noisy and perhaps short, and one does not know the dimension of the underlying system—let alone its governing equations. In this section, we turn our attention to that question using experimental data from two different dynamical systems: a far-infrared laser and a laboratory computer-performance experiment.

A Far-Infrared Laser

A canonical test case in the forecasting literature is the so-called "Dataset A" from the Santa Fe Institute prediction competition[16], which was gathered from a far-infrared laser. As in the synthetic examples in the previous section, the $\mathcal{A}_\tau$ and $MASE$ heatmaps (Figure 7) are largely antisymmetric for this signal. Again, there is a band across the bottom of each image because of the combined effects of overfolding and projection. Note the resemblance between Figures 7 and 5: the latter resemble "smoothed" versions of the former. It is well known[16] that the SFI A dataset is well described by the Lorenz 63 system with some added noise, so this similarity is both unsurprising and reassuring. LMA forecasts using the $\mathcal{A}_\tau$-optimal reconstruction of this trace were more accurate than similar forecasts using a reconstruction built using traditional heuristics ($MASE_{\mathcal{A}_\tau} = 0.0592$ versus $MASE_H = 0.0733$) and only slightly worse than the optimal value ($MASE_E = 0.0538$). However, the values of $\{m_{\mathcal{A}_\tau}, \tau_{\mathcal{A}_\tau}\}$ and $\{m_E, \tau_E\}$ are not identical for this signal. This is because the optima in the heatmaps in Figure 7 are bands, rather than unique points—as was the case in the synthetic examples in Section IV A. In a situation like this, a range of $\{m, \tau\}$ values are statistically indistinguishable, from the standpoint of the forecast accurary afforded by the corresponding reconstruction. The values suggested by the $\mathcal{A}_\tau$ calculation ($m_{\mathcal{A}_\tau} = 9$ and $\tau_{\mathcal{A}_\tau} = 1$) and by the exhaustive search ($m_E = 7$, $\tau_E = 1$) were all on this plateau[40]. Again, it appears that one can use $\mathcal{A}_\tau$ to choose good parameter values for delay reconstruction-based forecasting, but SFI A is only a single trace from a fairly simple system.



(a)



(b)

FIG. 7: The effects of reconstruction parameter values on $\mathcal{A}_\tau$ and forecast accuracy for "Dataset A" from the Santa Fe Institute time-series prediction competition. (a) $\mathcal{A}_\tau$ values for different delay reconstructions of SFI Dataset A. (b) $MASE$ scores for LMA forecasts on different delay reconstructions of SFI Dataset A.

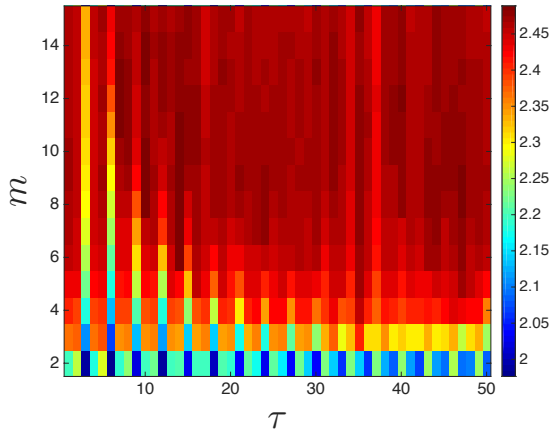
Computer Performance Dynamics

Laboratory experiments on computer performance dynamics have shown that these high-dimensional nonlinear systems exhibit a range of interesting deterministic dynamical behaviors[41, 42]. Both hardware and software play roles in these dynamics; changing either one can cause bifurcations from periodic orbits to low- and high-dimensional chaos. This rich range of behavior makes computer performance dynamics an ideal final test case for this paper.

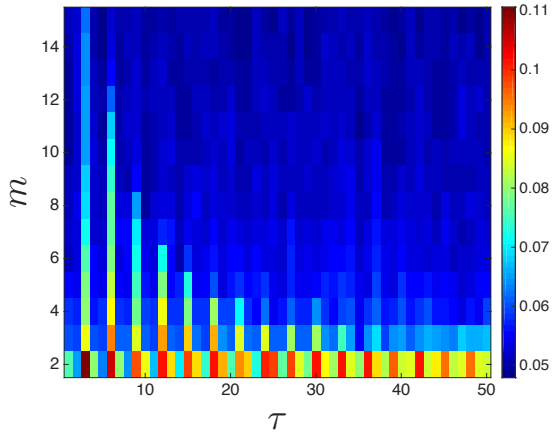
Collecting observations of the performance of a running computer requires some significant engineering. Basically, one programs the microprocessor's onboard hardware performance monitor to observe the quantities of interest, then stops the program execution at 100,000-instruction intervals—the unit of time in these experiments—and reads off the contents of those registers. Interested readers can find a detailed description of this custom measurement infrastructure in[42, 43]. The signals that are produced by this apparatus are scalar time-series measurements of system metrics like processor efficiency (*e.g.,* `IPC`, which measures how many instructions are being executed, on the average, in each clock cycle) or memory usage (*e.g.,* how often the processor had to access the main memory during the measurement interval).

Here, for conciseness, we focus on *processor* performance traces from two different programs, one simple and one complex, running on the same Intel i7-based computer. The first is four lines of C (`col_major`) that repeatedly initializes a 256×256 matrix in column-major order. The second is a much more complex program: the `403.gcc` compiler from the SPEC 2006CPU benchmark suite[44]. The performance traces of these two programs contained 147,925 points and 45,545 points, respectively. Since computer performance dynamics result from a composition of hardware and software, these two experiments involve two different dynamical systems, even though the programs are running on the same computer. But since other effects could be at work—housekeeping by the operating system, etc.—we repeated each experiment 15 times for a total of 30 traces. We have performed similar forecast experiments using other processor and memory performance metrics gathered during the execution of a variety of programs on several different computers [45]. Our preliminary analysis indicates that the results described in the rest of this section hold for those traces as well.

As in the previous examples, heatmaps of $MASE$ and $\mathcal{A}_\tau$ for the `col_major` time series (Figure 8b) are largely antisymmetric. And again, reconstructions using the $\mathcal{A}_\tau$-optimal parameter values allowed LMA to produce highly accurate forecasts of this signal: $MASE_{\mathcal{A}_\tau} = 0.0509$, compared to the optimal $MASE_E = 0.0496$. There are several major differences between these plots and the previous ones in this paper, though, beginning with the vertical stripes. These are due to the dominant unstable periodic orbit of period 3 in the chaotic attractor in the `col_major` dynamics. When τ is a multiple of this period ($\tau = 3\kappa$), the coordinates of the delay vector are not independent, which lowers $\mathcal{A}_\tau$ and makes forecasting more difficult. (There is a nice theoretical discussion of this effect in [3].) Conversely, $\mathcal{A}_\tau$ spikes and $MASE$ plummets when $\tau = 3\kappa - 1$, since the coordinates in such a delay vector cannot share any prime factors with the



(a)



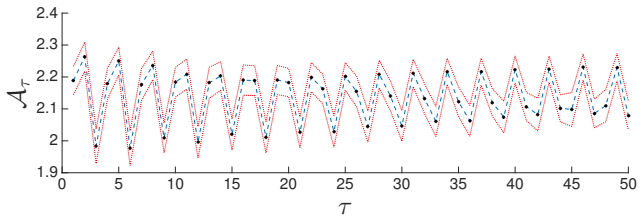
(b)

FIG. 8: The effects of reconstruction parameter values on $\mathcal{A}_\tau$ and forecast accuracy for a representative trace from a computer-performance dynamics experiment tracing the processor load during the execution of a simple program that repeatedly initializes a matrix in column-major order. (a) $\mathcal{A}_\tau$ values for different delay reconstructions of a `col_major` trace. (b) $MASE$ scores for LMA forecasts on different delay reconstructions of a `col_major` trace.

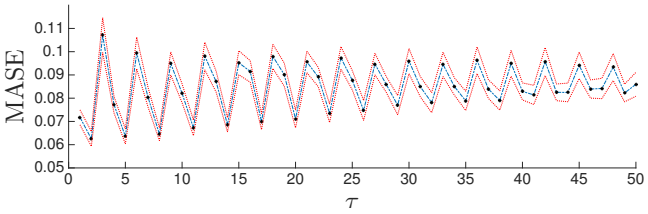
period of the orbit. The band along the bottom of both images is, again, due to a combination of overfolding and projection.

Another difference between the `col_major` heatmaps and the ones in Figures 4, 5, and 7 is the apparent overall trend: the "good" regions (low $MASE$ and high $\mathcal{A}_\tau$) are in the lower-left quadrants of those heatmaps, but in the upper-right quadrant of Figure 8. This is partly an artifact of the difference in the color-map scale, which was chosen here to bring out some important details of the structure, and partly due to that structure itself. Specifically, the optima of the `col_major` heatmaps—the large dark red and blue regions in Figures 8a and 8b, respectively—are much broader than the ones in the earlier sections of this paper, perhaps because the signal is so close to periodic. (This was also the case to some extent in the SFI A example, for the same reason.) This geometry makes precise comparisons of $\mathcal{A}_\tau$-optimal and $MASE$-optimal parameter values somewhat problematic, as the exact optima on two almost-flat but slightly noisy landscapes may not be in the same place. Indeed, the $\mathcal{A}_\tau$ values at $\{m_{\mathcal{A}_\tau}, \tau_{\mathcal{A}_\tau}\}$ and $\{m_E, \tau_E\}$ were within a standard error across all 15 traces of `col_major`.

And that brings up an interesting tradeoff. For practical purposes, what one wants is $\{m_{\mathcal{A}_\tau}, \tau_{\mathcal{A}_\tau}\}$ values that produce a $MASE$ value that is *close to* the optimum $MASE_E$. However, the algorithmic complexity of most nonlinear time-series analysis and prediction methods scales badly with m. In cases where the $\mathcal{A}_\tau$ maximum is broad, then, one might want to choose the lowest value of m on that plateau—or even a value that is on the *shoulder* of that plateau, if one needs to balance efficiency over accuracy. Indeed, forecasts with $m = 2$ appear to work surprisingly well for many nonlinear dynamical systems, including the `col_major` data[38]. Fixing $m = 2$ amounts to marginalizing the heatmaps in Figure 8, which produces a cross section like the ones shown in Figure 9. The



(a)



(b)

FIG. 9: $MASE$ and $\mathcal{A}_\tau$ for LMA forecasts of $m = 2$ delay reconstructions of all 15 `col_major` traces, plotted as a function of τ. The blue dashed curves show the averages across all trials; the red dotted lines are that average $\pm$ the standard deviation. (a) $\mathcal{A}_\tau$ values for delay reconstructions of the `col_major` traces with $m = 2$ and a range of values of τ. (b) $MASE$ scores for LMA forecasts of delay reconstructions of the `col_major` traces with $m = 2$ and a range of values of τ.

antisymmetry between $\mathcal{A}_\tau$ and $MASE$ is quite apparent in these plots; the global maximum of the former coincides with the global minimum of the latter, at $\tau = 2$. The average $MASE$ score of `col_major` forecasts con-

structed with $m = 2$ and this τ value is 0.0649. This is not much lower than the overall optimum of 0.0496—a value from a forecast whose free parameters required almost six orders of magnitude more CPU time to compute. As an important aside: these results suggest that one could bypass even more of the computational effort that is involved in delay reconstruction-based forecasting by simply working in two dimensions, i.e., by calculating $\mathcal{A}_\tau$ across a range of τs, rather than across a 2D $\{m, \tau\}$ space. This approach is discussed further in [38].

The correspondence between $MASE$ and $\mathcal{A}_\tau$ also holds true for other marginalizations: i.e., the minimum $MASE$ and the maximum $\mathcal{A}_\tau$ occur at the same τ value for all m-wise slices of the `col_major` heatmaps, to within statistical fluctuations. The methods of [5] and [14], incidentally, suggest $\tau_H = 2$ and $m_H = 12$ for these traces; the $MASE$ of an LMA forecast on such a reconstruction is 0.0530, which is somewhat better than the best result from the $m = 2$ marginalization, although still short of the overall optimum. The correspondence between τ_H and $\tau_{\mathcal{A}_\tau}$ is coincidence; for this particular signal, maximizing the independence of the coordinates happened to maximize the information about the future contained in each delay vector. The $m = 12$ result is not coincidence—and quite interesting, in view of the fact that the $m = 2$ forecast is so good. It is also surprising in view of the huge number of transistors—potential state variables—in a modern computer. As described in [42], however, the hardware and software constraints in these systems confine the dynamics to a much lower-dimensional manifold. All of these issues, and their relation to the task of prediction, are explored in more depth in [38].

The `col_major` program is what is known in the computer-performance literature as a "micro-kernel"—a extremely simple example that is used in proof-of-concept testing. The fact that its dynamics are so rich speaks to the complexity of the hardware (and the hardware-software interactions) in modern computers; again, see [42, 43] for a much deeper discussion of these issues. Modern computer programs are far more complex than this simple micro-kernel, of course, which begs the question: what does $\mathcal{A}_\tau$ tell us about the dynamics of truly complex systems like that—programs that the computer-performance community models as stochastic systems?

For `403.gcc`, the answer is, again, that $\mathcal{A}_\tau$ appears to be an effective and efficient way to assess predictability. It has been shown[45] that this time series shares little to no information with the future: i.e., that it *cannot* be predicted using delay reconstruction-based forecasting methods, regardless of τ and m values. The experiments in [45] required dozens of hours of CPU time to establish that conclusion; $\mathcal{A}_\tau$ gives the same results in a few seconds, using much less data. The structure of the heatmaps for this experiment, which are shown in Figure 10, is radically different. The patterns visible in the previous $MASE$ plots, and the antisymmetry between $\mathcal{A}_\tau$ and $MASE$ plots, are absent from Figure 10, reflecting the lack of predictive content in this signal. Note, too,

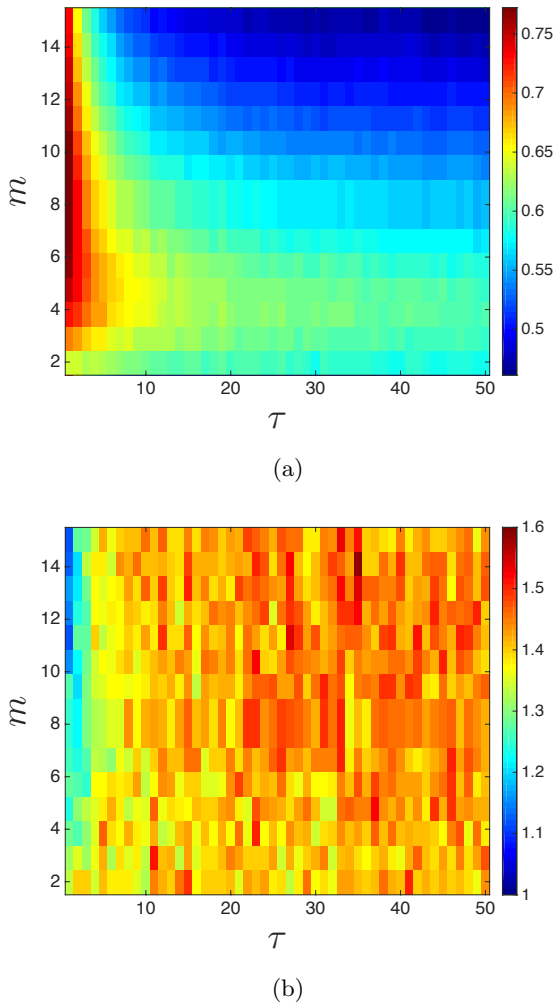


FIG. 10: The effects of reconstruction parameter values on $\mathcal{A}_\tau$ and forecast accuracy for a representative trace from a computer-performance dynamics experiment using the `403.gcc` benchmark. (a) $\mathcal{A}_\tau$ values for different delay reconstructions of a `403.gcc` trace. (b) $MASE$ scores for LMA forecasts on different delay reconstructions of a `403.gcc` trace.

that the color maps are different in this Figure. This reflects the much lower values of $\mathcal{A}_\tau$ for this signal: a maximum $\mathcal{A}_\tau$ of 0.7722 for `403.gcc`, compared to 5.3026 for Lorenz 96 with $K = 22$. Indeed, the $MASE$ surface in Figure 10b never dips below 1.0[46]. That is, regardless of parameter choice, LMA forecasts of `403.gcc` are no better than simply using the prior value of this scalar time series as the prediction. The uniformly low $\mathcal{A}_\tau$ values in Figure 10a are an effective indicator of this—and, again, they can be calculated quickly, from a relatively small sample of the data. It is to that issue that we turn next.

V. DATA REQUIREMENTS AND PREDICTION HORIZONS

In some real-world situations, it may be impractical to rebuild forecast models at every step, as we have done in the previous sections of this paper—because of computational expense, for instance, or because the data rate is very high. In these situations, one may wish to predict p time steps into the future, then stop and rebuild the model to incorporate the p points that have arrived during that period, and repeat. In chaotic systems, of course, there are fundamental limits on prediction horizon even if one is working with infinitely long traces of all state variables. A key question at issue in this section is how that effect plays out in forecast models that use delay reconstructions from scalar time-series data. And since real-world data sets are not infinitely long, it is important to understand the effects of data length on the estimation of $\mathcal{A}_\tau$.

A. Data Requirements for $\mathcal{A}_\tau$ Estimation

The quantity of data used in a delay reconstruction directly impacts the usefulness of that reconstruction. If one is interested in approximating the correlation dimension via the Grassberger-Procaccia algorithm, for instance, it has been shown that one needs $10^{(2+0.4m)}$ data points[47, 48]. Those bounds are overly pessimistic for forecasting, however. For example, Sugihara & May [20] used delay-coordinate reconstructions with m as large as seven to successfully forecast biological and epidemiological time-series data sets that contain as few as 266 points. A key challenge, then, is to determine whether one's time series *really* calls for as many dimensions and data points as the theoretical results require, or whether one can get away with fewer dimensions—and how much data one needs in order to figure all of that out.

We claim that $\mathcal{A}_\tau$ is a useful solution to those challenges. As established in the previous sections, calculations of this quantity can reveal what dimension one needs to build a good delay reconstruction for the purposes of LMA forecasting of nonlinear and chaotic systems. And, as alluded to in those sections, $\mathcal{A}_\tau$ can be estimated accurately from a surprisingly small number of points. The experiments in this section explore that intertwined pair of claims in more depth by increasing the length of the Lorenz 96 traces and testing whether the information content of the state estimator derived from standard heuristics converges to the $\mathcal{A}_\tau$-optimal estimator[49].

Figure 11 shows the results. When the data length is short, the $m = 2$ state estimator had the most information about the future. This makes perfect sense; a short time series cannot fully sample a complicated object, and when an ill-sampled high-dimensional manifold is projected into a low dimensional space, infrequently visited regions of that manifold can act effectively like noise. From an information-theoretic standpoint, this would increase the effective Shannon entropy rate of each of the variables in the delay vector. In the I-diagram in Figure 3, this would manifest as drifting apart of the two circles, decreasing the shaded region that one needs to maximize for effective forecasting.

If that reasoning is correct, longer data lengths should fill out the attractor, thereby mitigating the spurious increase in the Shannon entropy rate and allowing higher-dimensional reconstructions to outperform lower-dimensional ones. This is indeed what one sees in Figure 11. For both the $K = 22$ and $K = 47$ traces, once the signal is 2 million points long, the four-dimensional estimator has caught up to and even exceeded the two-dimensional case. Note, though, that the optimal $\mathcal{A}_\tau$ of the $m = 8$ reconstruction model is still lower than in the $m = 2$ or $m = 4$ cases, even at the right-hand limit of the plots in Figure 11. That is, even with a time series that contains 4×10^6 points, it is more effective to use a lower dimensional reconstruction to make an LMA forecast. But the really important message here is that $\mathcal{A}_\tau$ allows one to determine the best reconstruction parameters *for the available data*, which is an important part of the answer to the challenges outlined at the beginning of this section.

Something very interesting happens in the $m = 2$ results for Lorenz 96 model with $K = 47$: the $\mathcal{A}_\tau$ curve reaches a maximum value around 100,000 points and stops increasing, regardless of data length. What this means is that this two-dimensional reconstruction con-

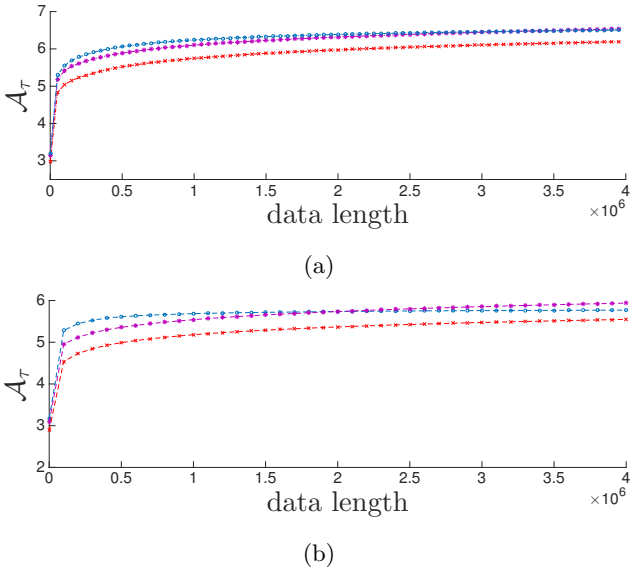


FIG. 11: $\mathcal{A}_\tau$ versus data length for traces from the Lorenz-96 system using $\tau = 1$ in all cases. Blue circles corresponds to an embedding dimension $m = 2$, purple diamonds to $m = 4$, and red xs to $m = 8$. (a) Optimal $\mathcal{A}_\tau$ for traces from the $\{K = 22, F = 5\}$ Lorenz 96 system. (b) Optimal $\mathcal{A}_\tau$ for traces from the $\{K = 47, F = 5\}$ Lorenz 96 system.

tains as much information about the future as can be ascertained from these data, suggesting that increasing the length of the training set would not improve forecast accuracy. To explore this, we constructed LMA forecasts of different-length traces (100,000–2.2 million points) from this system, then reconstructed their dynamics with different m values and the appropriate $\tau_{\mathcal{A}_\tau}$ for each case. For $m = 2$, both $\mathcal{A}_\tau$ and $MASE$ results did indeed plateau at 100,000 points—at 5.736 and 0.0809, respectively. As before, more data does afford higher-dimensional reconstructions more traction on the prediction problem: the $m = 4$ forecast accuracy surpassed $m = 2$ at around 2 million points ($MASE = 0.0521$). In neither case, by the way, did $m = 8$ catch up to either $m = 2$ or $m = 4$, even at 4 million data points. Of course, one must consider the cost of storing the additional variables in a higher-dimensional reconstruction model, particularly in data sets this long, so it may be worthwhile in practice to settle for the $m = 2$ forecast—which is only slightly less accurate and requires only 100,000 points. This has another major advantage as well. If the time series is non-stationary, a forecast strategy that requires fewer points can adapt more quickly.

B. Choosing reconstruction parameters for increased prediction horizons.

So far in this paper, we have considered forecasts that were constructed one step at a time and studied the correspondence of their accuracy with one-step-ahead calculations of $\mathcal{A}_\tau$. In this section, we consider longer prediction horizons (p) and explore whether one can use a p-step-ahead version of $\mathcal{A}_\tau$—i.e., $I[\mathcal{S}_j, X_{j+p}]$, with $p > 1$—to choose parameter values that maximize the information contained in each delay vector about the value of the time series p steps in the future.

One would expect the $\mathcal{A}_\tau$-optimal $\{m, \tau\}$ values for a given time series to depend on the prediction horizon. It has been shown, for instance, that longer-term forecasts generally do better with larger τ[6], and conversely[38]. It makes sense that one might need to reach different distances into the past (via the span of the delay vector) in order to reduce the uncertainty about events that are further into the future[16]. These effects are corroborated by $\mathcal{A}_\tau$. Figure 12 demonstrates this in the context of the Lorenz 96 system with $K = 22$, focusing on $m = 2$ for simplicity. The topmost trace in this figure is for the $p = 1$ case—i.e., a horizontal slice of Figure 4a made at $m = 2$. The maximum of this curve is the optimal τ value ($\tau_{\mathcal{A}_\tau}$) for this reconstruction. The overall shape of this trace reflects the monotonic increase in the uncertainty about the future with τ that is noted on page 5. The other traces in Figure 12 show $\mathcal{A}_\tau$ as a function of τ for $p = 2, 3, \ldots,$ down to $p = 100$ at the bottom of the figure. The lower traces do not decrease monotonically; rather, there is a slight initial rise. This is due to the point made above about the span of the delay vector: if one is predicting further into the future, it may be useful to reach further into the past. In general, this causes the optimal τ to shift to the right as prediction horizon increases, going down the plot—i.e., longer prediction horizons require a greater τ (cf. [6]). For very long horizons, the choice of τ appears to matter very little. In particular, $\mathcal{A}_\tau$ is fairly constant and quite low for $5 < \tau < 50$ when $p > 30$—i.e., regardless of the choice of τ, there is very little information about the p-distant future in any delay reconstruction of this signal for $p > 30$. This effect should not be surprising, and it is well corroborated in the literature. However, it can be hard to know *a priori*, when one is confronted with a data set from an unknown system, to know what prediction horizon makes sense. $\mathcal{A}_\tau$ offers a computationally efficient way to answer that question.

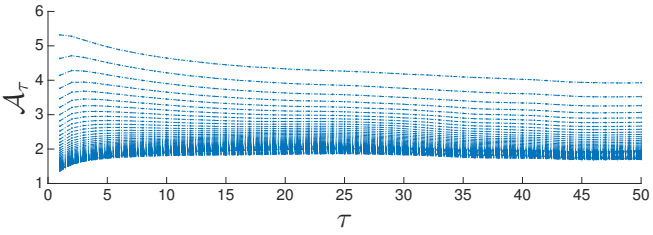


FIG. 12: The effect of prediction horizon (p) on $\mathcal{A}_\tau$ of the $K = 22$ Lorenz 96 system for a fixed reconstruction dimension ($m = 2$). The traces in the image, starting from the top, correspond to prediction horizons of $p = 1$ to $p = 100$.

Figure 13 shows a similar exploration of the other side of that question: the effects of the reconstruction dimension on $\mathcal{A}_\tau$, with τ fixed at 1. The $m = 2$ state estimator contains more information about the future for short prediction horizons. This ties back to the discussion at the end of Section IV B: low-dimensional reconstructions can work quite well for short prediction horizons. However, Figure 13 shows that the full reconstruction is better for longer horizons. This is not terribly surprising, since a higher reconstruction dimension allows the state

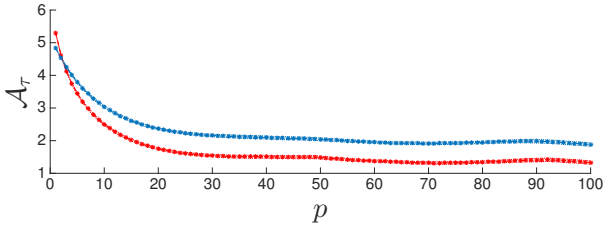


FIG. 13: The effect of prediction horizon (p) on $\mathcal{A}_\tau$ of the $K = 22$ Lorenz 96 system for a fixed time delay ($\tau = 1$) and two different reconstruction dimensions. The red line is $m = 2$ and the blue is $m_H = 8$, the value suggested for this signal by the technique of false neighbors.

estimator to capture more information about the past. Finally, note that $\mathcal{A}_\tau$ decreases monotonically with prediction horizon for both $m = 2$ and m_H. This, too, is unsurprising. Pesin's relation[50] says that the sum of the positive Lyapunov exponents is equal to the entropy rate, and if there is a non-zero entropy rate, then generically observations will become increasingly independent the further apart they are. This explanation also applies to Figure 12, of course, but it does *not* hold for signals that are wholly (or nearly) periodic.

Recall that the `col_major` dynamics in Section IV B were chaotic, but with a dominant unstable periodic orbit—which had a variety of interesting effects in the results. Figure 14 explores the effects of prediction horizon on those results. Not surprisingly, there is some periodicity in the $\mathcal{A}_\tau$ versus p relationships, but not for the same reasons that caused the stripes in Figure 8b. Here, the *peaks* in $\mathcal{A}_\tau$ occur at multiples of the period. That is, the $m = 2$ state estimator can forecast with the most success when the value being predicted is in phase with the delay vector. Note that this effect is far stronger for $m = 2$ than m_H, simply because of the instability of that periodic orbit; the visits made by the chaotic trajectory to that orbit are more likely to be short than long. As expected, $\mathcal{A}_\tau$ decays with prediction horizon—but only at first, after which it begins to rise again, peaking at $p = 69$ and $p = 71$. This may be due to a second higher-order unstable periodic orbit in the `col_major` dynamics.

In theory, one can derive rigorous bounds on prediction horizon. The time at which $\mathcal{S}_j$ will no longer have any information about the future can be determined by considering:

$$R(p) = \frac{I[\mathcal{S}_j; X_{j+p}]}{H[X_{j+p}]} \ ,$$

i.e., the percentage of the uncertainty in X_{j+p} that can be reduced by the delay vector. Generically, this will limit to some small value equal to the amount of information that the delay vector contains about any arbitrary point on the attractor. Given some criteria regarding how much information above the "background" is required of the state estimator, one could use the $R(p)$ versus p curve to determine the maximum practical horizon.

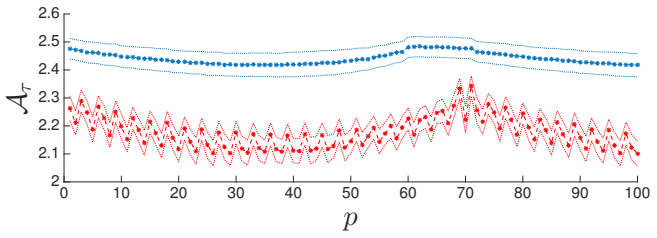


FIG. 14: The effect of prediction horizon (p) on $\mathcal{A}_\tau$ of the `col_major` for a fixed time delay ($\tau = 1$) and two different reconstruction dimensions. The red line is $m = 2$ and the blue is $m_H = 12$, the value suggested for this signal by the technique of false neighbors.

In practice, one can select parameters for delay reconstruction-based forecasting by explicitly including the prediction horizon in the $\mathcal{A}_\tau$ function, fixing the horizon p at the required value, performing the same search as we did in earlier sections over a range of m and τ, and then choosing a point on (or near) the optimum of that $\mathcal{A}_\tau$ surface. The computational and data requirements of this calculation, as shown in Section V A, are far superior to those of the standard heuristics used in delay reconstructions.

VI. CONCLUSION

In this paper, we have described a new metric for quantifying how much information about the future is contained in a delay reconstruction. Using a number of different dynamical systems, we demonstrated a direct correspondence between the $\mathcal{A}_\tau$ value for different delay reconstructions and the accuracy of forecasts made with Lorenz's method of analogues on those reconstructions. Since $\mathcal{A}_\tau$ can be calculated quickly and reliably from a relatively small amount of data, without needing to know anything about the governing equations or the state space dynamics of the system, that correspondence is a major advantage, in that it allows one to choose parameter values for delay reconstruction-based forecast models without doing an exhaustive search on the parameter space. Significantly, $\mathcal{A}_\tau$-optimal reconstructions are better, for the purposes of forecasting, than reconstructions constructed using standard heuristics like mutual information and the method of false neighbors, which can require large amounts of data, significant computational effort, and expert human interpretation. $\mathcal{A}_\tau$ allows one to answer other questions regarding forecasting with theoreticaly unsound models[38]—e.g., why one can obtain a better forecast using a low-dimensional reconstruction than with a full embedding. It also allows one to understand bounds on prediction horizon without having to estimate Lyapunov spectra or Shannon entropy rates, which are difficult to obtain for arbitrary real-valued time series. That, in turn, allows one to tailor one's reconstruction parameters to the amount of available data and the desired prediction horizon—and to know if a given prediction task is just not possible.

The explorations in this paper involve a simple near-neighbor forecast strategy and state estimators that are basic delay reconstructions of raw time-series data. The definition and calculation of $\mathcal{A}_\tau$ do not involve any assumptions about the state estimator, though, so the results presented here should also hold for other state estimators. For example, it is common in time-series prediction to pre-process one's data: for example, low-pass filtering or interpolating to produce additional points. Calculating $\mathcal{A}_\tau$ after performing such an operation will accurately reflect the amount of information in that new time series—indeed, it would reveal if that pre-processing

step *destroyed* information. And we believe that the basic conclusions in this paper extend to other state-space based forecast schemas besides Lorenz's method of analogues, such as those used in [16–18, 20, 28]—although $\mathcal{A}_\tau$ may not accurately select optimal parameter values for strategies that involve post-processing the data (e.g., GHKSS[51]). We are in the process of exploring this.

There are many other interesting potential ways to leverage $\mathcal{A}_\tau$ in the practice of forecasting. If the $\mathcal{A}_\tau$-optimal $\tau = 1$, that may be a signal that the time series is undersampling the dynamics and that one should increase the sample rate. One could use the more general form $\mathcal{A}_\mathcal{S}$ at a finer grain to optimizing τ individually for each dimension, as suggested in [52–54] where optimal optimal values are selected based on criteria not directly related to prediction. To do this, one could define $\mathcal{S}_j = [X_j, X_{j-\tau_1}, X_{j-\tau_2}, \ldots, X_{j-\tau_{m-1}}]$ and then simply maximize $\mathcal{A}_\mathcal{S}$ using that state estimator constrained over $\{\tau_i\}_{i=1}^{m-1}$.

ACKNOWLEDGEMENTS

We thank Jim Crutchfield for many helpful discussions. This material is based upon work supported by, or in part by, NSF Grant No. CMMI-1162440, and the U. S. Army Research Laboratory and the U. S. Army Research Office under contracts W911NF-13-1-0390 and W911NF-13-1-0340.

VII. REFERENCES

[1] F. Takens. Detecting strange attractors in fluid turbulence. In D. Rand and L.-S. Young, editors, *Dynamical Systems and Turbulence*, pages 366–381. Springer, Berlin, 1981.

[2] N. Packard, J. Crutchfield, J. Farmer, and R. Shaw. Geometry from a time series. *Physical Review Letters*, 45:712, 1980.

[3] T. Sauer, J. Yorke, and M. Casdagli. Embedology. *Journal of Statistical Physics*, 65:579–616, 1991.

[4] E. Olbrich and H. Kantz. Inferring chaotic dynamics from time-series: on which length scale determinism becomes visible. *Phys. Lett. A*, 232(1-2):63–69, 1997.

[5] A. Fraser and H. Swinney. Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33(2):1134–1140, 1986.

[6] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge, 1997.

[7] Th. Buzug and G. Pfister. Comparison of algorithms calculating optimal embedding parameters for delay time coordinates. *Physica D: Nonlinear Phenomena*, 58(1-4):127 – 137, 1992.

[8] W. Liebert, K. Pawelzik, and H. Schuster. Optimal embeddings of chaotic attractors from topological considerations. *Europhysics Letters*, 14(6):521–526, 1991.

[9] Th. Buzug and G. Pfister. Optimal delay time and embedding dimension for delay-time coordinates by analysis of the global static and local dynamical behavior of strange attractors. *Physical Review A*, 45:7073–7084, May 1992.

[10] W. Liebert and H. Schuster. Proper choice of the time delay for the analysis of chaotic time series. *Physics Letters A*, 142(2-3):107 – 111, 1989.

[11] M. Rosenstein, J. Collins, and C. De Luca. Reconstruction expansion as a geometry-based framework for choosing proper delay times. *Physica D: Nonlinear Phenomena*, 73(1-2):82–98, May 1994.

[12] L. Cao. Practical method for determining the minimum embedding dimension of a scalar time series. *Physica D: Nonlinear Phenomena*, 110(1-2):43–50, 1997.

[13] D. Kugiumtzis. State space reconstruction parameters in the analysis of chaotic time series—The role of the time window length. *Physica D: Nonlinear Phenomena*, 95(1):13–28, 1996.

[14] M. Kennel, R. Brown, and H. Abarbanel. Determining minimum embedding dimension using a geometrical construction. *Physical Review A*, 45:3403–3411, 1992.

[15] R. Hegger, H. Kantz, and T. Schreiber. Practical implementation of nonlinear time series methods: The TISEAN package. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 9(2):413–435, 1999.

[16] A. Weigend and N. Gershenfeld, editors. *Time Series Prediction: Forecasting the Future and Understanding the Past*. Santa Fe Institute Studies in the Sciences of Complexity, Santa Fe, NM, 1993.

[17] M. Casdagli and S. Eubank, editors. *Nonlinear Modeling and Forecasting*. Addison Wesley, 1992.

[18] L. Smith. Identification and prediction of low dimensional dynamics. *Physica D: Nonlinear Phenomena*, 58(1–4):50 – 76, 1992.

[19] A. Pikovsky. Noise filtering in the discrete time dynamical systems. *Soviet Journal of Communications Technology and Electronics*, 31(5):911–914, 1986.

[20] G. Sugihara and R. May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 344:734–741, 1990.

[21] E. Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences*, 26:636–646, 1969.

[22] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 13(1):25–54, 2003.

[23] R. W Yeung. *A first course in information theory*. Springer Science & Business Media, 2012.

[24] R. G. James, J. R. Mahoney, C. J. Ellison, and J. P. Crutchfield. Many roads to synchrony: Natural time scales and their algorithms. *Physical Review E*, 89(4):042135, 2014.

[25] C. C. Strelioff and J. P. Crutchfield. Bayesian structural inference for hidden processes. *Physical Review E*, 89(4):042119, 2014.

[26] A. J. Bell. The co-information lattice. In *Proceedings of 4th International Symposium on Independent Component Analysis and Blind Source Separation*, pages 921–926, 2003.

[27] H. W. Sorenson. *Kalman Filtering: Theory and Application*. IEEE Press, 1985.

[28] T. Sauer. Time-series prediction by using delay-coordinate embedding. In *Time Series Prediction: Forecasting the Future and Understanding the Past*. Santa Fe Institute Studies in the Sciences of Complexity, Santa Fe, NM, 1993.

[29] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[30] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000.

[31] S. Frenzel and B. Pompe. Partial mutual information for coupling analysis of multivariate time series. *Phys. Rev. Lett.*, 99:204101, Nov 2007.

[32] J. T. Lizier. JIDT: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI*, 1(11), 2014.

[33] The formula for the other KSG estimation algorithm is subtly different; it sets r_x and r_y to the maxima of the x and y distances to the k nearest neighbors.

[34] R. Hyndman and A. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.

[35] E. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20:130–141, 1963.

[36] E. Lorenz. Predictability: A problem partly solved. In T. Palmer and R. Hagedorn, editors, *Predictability of Weather and Climate*, pages 40–58. Cambridge University Press, 2006.

[37] A. Karimi and M. Paul. Extensive chaos in the Lorenz-96 model. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(4), 2010.

[38] J. Garland and E. Bradley. Prediction in projection. *Chaos*, 25:123108, 2015.

[39] Note that the color map scales are not identical across all heatmap figures in this paper; rather, they are chosen individually, to bring out the details of the structure of each experiment.

[40] The values suggested by the traditional heuristics, $m_H = 7$ and $\tau_H = 3$, were off the shoulder of that plateau.

[41] Z. Alexander, T. Mytkowicz, A. Diwan, and E. Bradley. Measurement and dynamical analysis of computer performance data. In *Proceedings of Advances in Intelligent Data Analysis IX*, volume 6065. Springer Lecture Notes in Computer Science, 2010.

[42] T. Myktowicz, A. Diwan, and E. Bradley. Computers are dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 19(3), 2009.

[43] T. Mytkowicz. *Supporting experiments in computer systems research*. PhD thesis, University of Colorado, November 2010.

[44] J. Henning. SPEC CPU2006 benchmark descriptions. *SIGARCH Computer Architecture News*, 34(4):1–17, 2006.

[45] J. Garland, R. James, and E. Bradley. Model-free quantification of time-series predictability. *Physical Review E*, 90:052910, 2014.

[46] Figure 4, in contrast, never exceeds ≈ 0.6 and generally stays below 0.3.

[47] A. A. Tsonis, J. B. Elsner, and K. P. Georgakakos. Estimating the dimension of weather and climate attractors: Important issues about the procedure and interpretation. *Journal of the Atmospheric Sciences*, 50(15):2549–2555, 1993.

[48] L. Smith. Intrinsic limits on dimension calculations. *Physical Letters A*, 133(6):283–288, 1988.

[49] This kind of experiment is not possible in practice, of course, when the time series is fixed, but can be done in the context of this synthetic example.

[50] Y. B. Pesin. Characteristic lyapunov exponents and smooth ergodic theory. *Russian Mathematical Surveys*, 32(4):55–114, 1977.

[51] P. Grassberger, R. Hegger, H. Kantz, C. Schaffrath, and T. Schreiber. On noise reduction methods for chaotic data. *Chaos*, 3:127, 1993.

[52] L. Pecora, L. Moniz, J. Nichols, and T. Carroll. A unified approach to attractor reconstruction. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 17(1), 2007.

[53] M. Small and C.K. Tse. Optimal embedding parameters: a modelling paradigm. *Physica D: Nonlinear Phenomena*, 194(3–4):283 – 296, 2004.

[54] C. Nichkawde. Optimal state-space reconstruction using derivatives on projected manifold. *Phys. Rev. E*, 87:022905, Feb 2013.