Nonlinear Dyn (2022) 108:2147–2160
https://doi.org/10.1007/s11071-022-07311-y

ORIGINAL PAPER

Empirical mode modeling

A data-driven approach to recover and forecast nonlinear dynamics from noisy data

Joseph Park · Gerald M. Pao · George Sugihara · Erik Stabenau · Thomas Lorimer

Received: 13 October 2020 / Accepted: 17 February 2022 / Published online: 18 March 2022

Abstract Data-driven, model-free analytics are natural choices for discovery and forecasting of complex, nonlinear systems. Methods that operate in the system state-space require either an explicit multidimensional state-space, or, one approximated from available observations. Since observational data are frequently sampled with noise, it is possible that noise can corrupt the state-space representation degrading analytical performance. Here, we evaluate the synthesis of empirical mode decomposition with empirical dynamic modeling, which we term empirical mode modeling, to increase the information content of state-space representations in the presence of noise. Evaluation of a mathematical, and, an ecologically important geophysical application across three different state-space representations suggests that empirical mode modeling may be a useful technique for data-driven, model-free, state-space analysis in the presence of noise.

Keywords Empirical mode decomposition · Empirical dynamic modeling · Empirical mode modeling · Data-driven analysis · Nonlinear systems

J. Park (✉)
United Nations Comprehensive Nuclear-Test-Ban Treaty Organization, Department of Engineering and Development, Vienna, Austria
e-mail: JosephPark@IEEE.org

J. Park · E. Stabenau
U.S. Department of the Interior, South Florida Natural Resources Center, Homestead, FL 33031, USA

G. M. Pao
Salk Institute for Biological Studies, MCBL-4, La Jolla, CA 92037, USA
e-mail: pao@Salk.edu

G. M. Pao
Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna-son, Kunigami-gun Okinawa 904-0495, Japan
e-mail: GERALD.PAO@OIST.JP

G. Sugihara · T. Lorimer
Scripps Institution of Oceanography Organization University of California San Diego, La Jolla, CA 92037, USA

1 Introduction

Evolution of science and technology is episodically redirected as our understanding improves. For example, the late 19th to mid-20th Century recognition of nonlinear dynamical systems as not purely stochastic, or, purely deterministic was both troubling and opportune. The subsequent acknowledgment of emergent behaviors as a *property* of such systems can be considered such a redirection[1]. Progress along this front established dynamical systems theory as a new branch of science [2]. Accordingly, it is now well-accepted that canonical statistical and parametric equation-based

models applied to nonlinear systems, which turn out to be nearly ubiquitous in complex technologies and nature, are often problematic [3].

Currently, machine learning is advancing new directions in data and system analysis. Techniques such as manifold learning [4] and diffusion maps [5] rely on a *data-driven*, inductive approach, wherein the data-itself reveals underlying dynamics and behavioral complexities, rather than a presumptive, model-based deductive analysis. Empirical dynamic modeling (EDM) is one such paradigm, having developed into a useful toolset for system analysis and forecasting [6–10].

EDM operates in multidimensional state-space, either from an explicit multidimensional data set, or, from a diffeomorphic reconstruction of the state-space, for example, by application of Takens embedding theorem [11,12]. EDM is not based on parametric presumptions, fitting statistics, or, specifying equations; representing a data-driven approach amenable to nonlinear dynamics as noted by DeAngelis [3]. Particular strengths of EDM are time series forecasting, and, identification of intervariable interactions and causal relationships between system variables. The unfamiliar reader is referred to the lucid EDM introduction in reference [13].

Since these methods are data-driven, it is possible that observational noise can interfere with accurate state-space representations, degrading inference and forecast skill. We therefore seek methods to improve state-space representations derived from noisy observations, and here, turn to empirical mode decomposition (EMD).

EMD decomposes oscillatory signals into scale-dependent modes termed intrinsic mode functions (IMF) without constraints of linearity or stationarity as presumed by Fourier, wavelet or eigendecomposition. Application of the Hilbert transform to IMFs provides time-dependent instantaneous frequency estimates, with the combination of EMD and IMF Hilbert spectra constituting the Hilbert-Huang transform (HHT) [14]. IMFs are particularly astute at isolating physically-meaningful dynamics [14], and, their data-driven isolation of noise is a cornerstone of EMD utility [15]. An introduction and review are found in references [14] and [16].

Since dynamics of interest are often low-dimensional and contained within a bounded state-space, univariate projections of state-space trajectories are often expressed as oscillatory signals. EMD provides a model-free and natural way to filter noise and isolate physically relevant modes of oscillatory time series, while EDM enables data-driven state-space metrics for forecasting and quantifying cross-variable interactions.

Here, we combine the model-free modal decomposition of EMD with the multivariate state-space representation inherent in EDM to improve forecasting and discovery in the presence of noisy or confounded observations. We term this empirical mode modeling (EMM).

We demonstrate the utility of this synthesis in two distinct applications. First a nonlinear, chaotic mathematical system, the Rössler attractor, and second, a nonlinear geophysical system with ecological importance, salinities in Florida Bay within Everglades National Park.

1.1 Model fitness and application

Following Granger [17], we adopt a position that model predictability reflects the information a model contains regarding a dynamical system. We therefore use model fidelity expressed as the Pearson correlation ρ between observations and model predictions as a fitness metric. Note that the models themselves can be equation-free and nonlinear, and, that other metrics such as mutual information are equally valid.

When models are used in an operational forecast system it is typical to assess model fitness on *out-of-sample* data, that is, the model is trained on a subset of available data with fitness assessed on a prediction (validation) set disjoint from the training data. Out-of-sample prediction characterizes model stability and the ability to generalize to untrained data.

In-sample models, where the training and prediction sets overlap, are informative when discovering/assessing the ability of the model to represent system dynamics. For example, as a function of state-space variables and their interactions, or, in response to noise/confounded states. We use both formats to assess model fitness and dynamical information content.

1.2 Model representation and alternatives

The application of EMM presumes the state-space is adequately represented. When approximating the state-space from time series observations the length of data

as well as it's sampling interval enforce constraints on state-space fidelity to the underlying dynamics [18]. Munch et al. [19] examine these issues with specific relevance to EDM.

Another key parameter is the dimension of the reconstructed state-space. The choice of the "correct" dimension is fundamental, as an underestimated dimension in the reconstruction may not adequately resolve the state-space. This can result in the appearance of a random component in the lower-dimensional dynamics, even when the data is purely deterministic. As noted below, an initial step in EDM is to determine the dimensionality that best-represents the data in the state-space reconstruction. The reader is referred to Chang et al.[13] for an introduction and demonstration. This is exemplified in the analysis of hypersalinity in Sect. 4, shown explicitly in "Appendix 1" (Sect. 1).

We also note that the synthesis of EMD with other operators can provide an alternative to EDM for the dynamical system analysis. For example, a stochastic evolution operator may serve to alleviate constraints on data representation [20,21].

2 Empirical mode modeling

The basis of EMM is to create a multidimensional state-space from empirical mode decomposition of observational time series. The resultant IMFs, or subsets of IMFs representing the system state-space, are then analyzed within the empirical dynamic modeling framework. The reader is referred to reference [16] for details of empirical mode decomposition, and reference [13] for empirical dynamic modeling.

Generically, EMM can be implemented with the following steps:

1. Determine the signal-to-noise ratio (SNR) of the time series. If the SNR is less than 3 dB, EMM may provide improvements in state-space representation and forecasting.
2. Apply empirical mode decomposition to the time series creating a set of intrinsic mode functions (IMFs). This can be done for multiple time series.
3. Optionally select IMFs that best represent the dynamics. This is an ongoing area of research. Here, we use two methods:
   (a) Multiview embedding as described in reference [10], selecting the set of IMFs that maximize model predictability as exemplified in Sect. 3.2.2.
   (b) Select IMFs based on spectral filtering, or, manual inspection to remove high frequency noise or low frequency oscillations external to the dynamics. This is shown in Sect. 4.1.
4. Apply empirical dynamic modeling using the selected IMFs as a multivariable state-space. Application of EDM to multivariable state-space is described in reference [8]. EDM analysis can be performed using simplex [6], sequential locally weighted global linear maps (S-maps) [7], or convergent cross mapping [9].

3 Rössler dynamics

The Rössler attractor is a three-dimensional coupled dynamical system known to exhibit chaotic dynamics [22]. Our implementation is defined as:

$$\frac{dx}{dt} = -y - z \tag{1}$$

$$\frac{dy}{dt} = x + ay \tag{2}$$

$$\frac{dz}{dt} = b + z(x - c) \tag{3}$$

with constants $a = 0.4$, $b = 0.4$, $c = 4$, initial state values $x_0 = 1$, $y_0 = 0$, $z_0 = 1$, time increment = 0.01, and time span T = [0-500]. We then subsample the integrated solutions of equations 1 - 3 at a time increment of 0.1, and ignore the time span from 0 to 200 resulting in a 3-D time series of 3000 points from time 200 to 500. Multispectral noise (N) is generated following reference [23] as a combination of brown and pink noise:

$$N = A\ (B\ p_n + C\ b_n) \tag{4}$$

Table 1 Rössler signal-to-noise ratios

Noise A	1	2	4	8	12	16	24	32	48	64
SNR dB	10.08	7.00	4.12	1.07	–0.73	–1.89	–3.71	–4.94	–6.69	–7.95

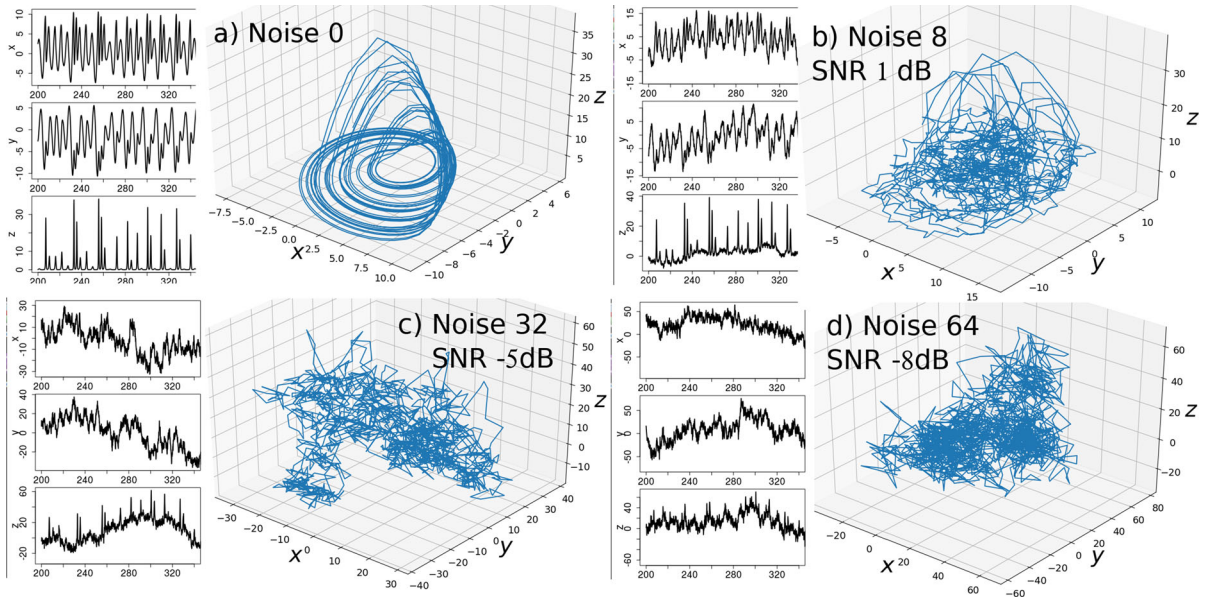

Fig. 1 State-space time series and trajectories of Rössler dynamics with multispectral noise

where p_n is "pink" noise with power spectral density $1/f$, b_n is low frequency "Brownian" noise with power spectrum $1/f^2$, where f is frequency, and, noise amplitude A ranges from 0 to 64; $B = 0.5$, $C = 1$. Corresponding signal-to-noise ratios are listed in Table 1.

Figure 1 shows four examples of the Rössler state-space with different levels of additive noise. It is clear that as noise increases, trajectories in the state-space become corrupted and entangled, and, we expect that methods such as manifold learning and EDM that rely on state-space representations of system dynamics can perform poorly as noise increases.

3.1 Empirical mode decomposition

The first step in EMM is to obtain IMFs of the observation time series. Figure 2 shows the Rössler variable x and it's empirical mode decomposition IMFs at four values of additive multispectral noise amplitude A from equation 4. We can see that IMFs partition timescale variance into modes with decreasing ranges of instantaneous frequency, high frequencies in the low-order modes and lower frequencies in the higher-order modes. We will leverage this partition as a data-driven filter to select IMFs that better represent the underlying dynamics.

3.2 Empirical dynamic modeling

Empirical dynamic modeling provides state-space forecasting and cross-variable analysis, here, we use the simplex algorithm to project univariate state-space dynamics to assess the fitness of state-space representations derived from traditional time-delay embeddings, and, from IMFs of observed variables. IMFs are computed over the entire time series, and, partitioned into training library and prediction sets in the same manner as any EDM candidate time series. The forecast variable is z from Eq. 3, with forecasts made from three different state-space representations consisting of vectors $[x, y]$, $[x, y, z]$, or their IMFs.

The simplex algorithm is a state-space projection from a simplex of nearest neighbors in the training data (*library*) closest to the state-space of the prediction point. The simplex consists of $E + 1$ points, where E is the embedding dimension in the case the state-space is a time-delay embedding, or, simply the dimension of the multivariate state-space. Reference [6] describes the simplex algorithm.

3.2.1 Model comparison

We examine four different EDM models (Table 2). The first is a full variable state-space of $[x, y, z]$ $(E = 3)$

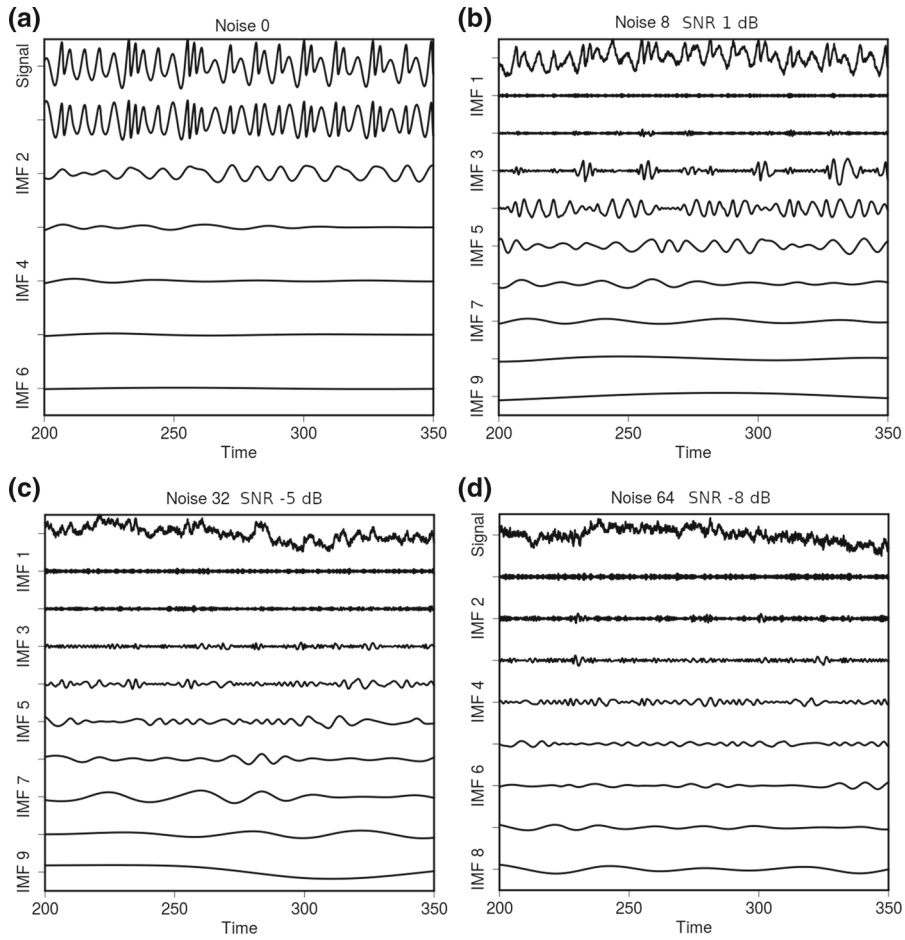

Fig. 2 Rössler state-space variable x time series with 3 instances of additive noise, and the resultant IMFs

Table 2 Rössler EDM State-space Variables

State-space	Variables	E
Multivariable	x, y, z	3
Multivariable	x, y	2
Takens time-delay	x, y	6
Multivariable	IMF_x, IMF_y	N_{IMF} (12-18)

with complete information. This serves as a reference. Second, a multivariable $[x, y]$ $(E = 2)$ state-space representing a naive predictor without information of the forecast variable z. Third, a Takens time-delay state-space of variables $[x, y]$ with embedding dimension $E_e = 3$ resulting in a state-space dimension of $E = 6$, and fourth, multivariate state-space of all IMFs of $[x, y]$. Here, the dimension E is equal to the number of IMFs.

Since we are interested in the ability of these different representations to model the underlying dynamics, all models use the target time series of noiseless variable z, and perform simplex projection at a forecast interval of $T_p = 0$. All models are evaluated in an out-of-sample prediction with the training library consisting of points [1-2000], and prediction set [2001-3000].

Figure 3 shows ensemble simplex projection results as function of signal-to-noise ratio for the four state-space representations over 1000 noise realisations at each noise amplitude. Here, we find the naive predictor $x, y \rightarrow z$ incapable of meaningful prediction, while the Takens time-delay embedded model (Takens $x, y \rightarrow z$) approaches that of the reference model $(x, y, z \rightarrow z)$ at high signal-to-noise ratios. At high SNR, above approximately 3 dB, the Takens model outperforms EMM based on IMFs, however, as SNR falls below 3 dB EMM outperforms the time-delay representation providing information recovery closer to the full-information reference model. A SNR of 3 dB translates to signal power a factor of 2 greater than the noise power.

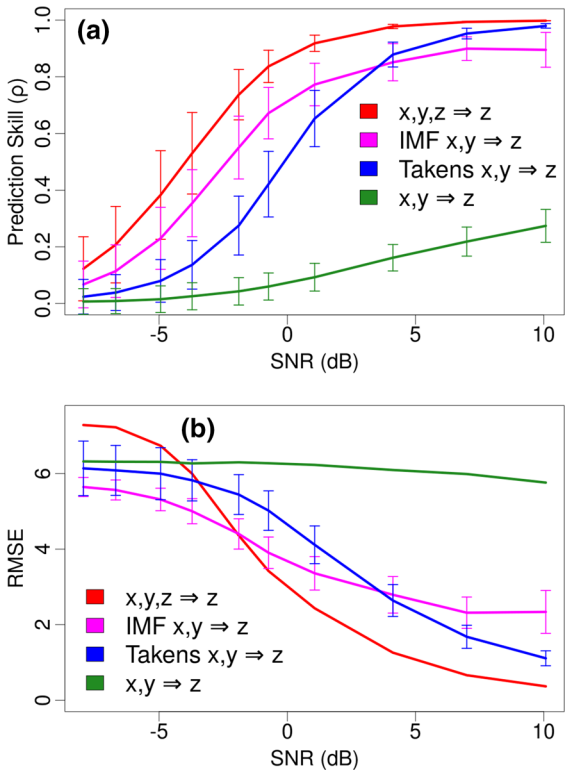

Fig. 3 Pearson correlation and RMSE of simplex projected Rössler variable z as a function of noise over 1000 noise realisations at each noise amplitude. Lines are mean values, error bars ± standard error

As noted, at SNR greater than 3 dB, EMM performs slightly worse than the time-delay embedding. It may be that the lower dimensional time-delay embedding provides a more compact state-space simplex that better represents state evolution when noise is low. Here, the time-delay state-space is 6 dimensional, while the EMM multivariate space is 12 dimensional.

An alternative EDM algorithm: Sequential Locally Weighted Global Linear Maps (S-map) [7] applies a localisation kernel to the library of states, rather than using a fixed number of k-nearest-neighbors to define a simplex. If dimensional inflation of the simplex is the cause of degraded performance at high SNR, S-map based EMM may address this issue. However, the goal of EMM is to improve state-space models in the presence of noise.

These results suggest that for the data examined, when noise power is more than 1/2 the signal power, even an indiscriminate use of IMFs as EDM state-space variables instead of time-delay embedding can improve cross-variable projection $x, y \rightarrow z$, evincing a more inforamtic representation of the underlying dynamics.

3.2.2 IMF selection

It was observed that low order IMFs capture high frequency noise not present in the underlying Rössler dynamics, and, that use of all IMFs as EDM state-space variables improves predictability in the presence of noise. It is then natural to ask whether subsets of IMFs can provide additional gains in state-space representation and model fidelity, or, provide equitable predictability from a lower-dimensional representation.

One method to identify noise-dominated IMFs is to examine the variance of IMF instantaneous frequencies (IF) since high variance in instantaneous frequency is indicative of noise, while low variance suggests a relatively stable oscillatory mode reflective of low-dimensional dynamics. The idea being to discard IMFs with instantaneous frequency variance above some threshold. However, this does not set a useful bound on low frequency IMFs, and requires a data-specific threshold selection.

Following the model agnostic, data-driven approach wherein model predictability serves as a metric of information content, Ye & Sugihara suggested multiview embedding as an EDM algorithm to select maximally informative state-space variables [10]. Multiview embedding evaluates model fidelity across all combinations of a set of candidate variables, selecting the D-dimensional subset with the highest predictive skill. In standard multiview embedding, the selection process is performed with in-sample simplex projection, the top ranked variables used in an out-of-sample prediction for the final model output. However, in-sample ranking poses a problem with IMFs since low-frequency IMFs approach, and eventually express monotonic functions.

Simplex projection is a nonlinear state-space mapping from state-space variables to a target without constraint on the variables. If the variables support a unique mapping across their domain, then good *in-sample* predictability can be achieved. This means that arbitrary, non-constant, non-oscillatory functions that provide a unique multivariable mapping can be mapped in-sample with good fidelity to the target. Since low frequency IMFs have very little or no oscillatory content, they are likely to be selected as in-sample variables with high predictive skill. Normally, candidate state-space variables are oscillatory or physically relevant,

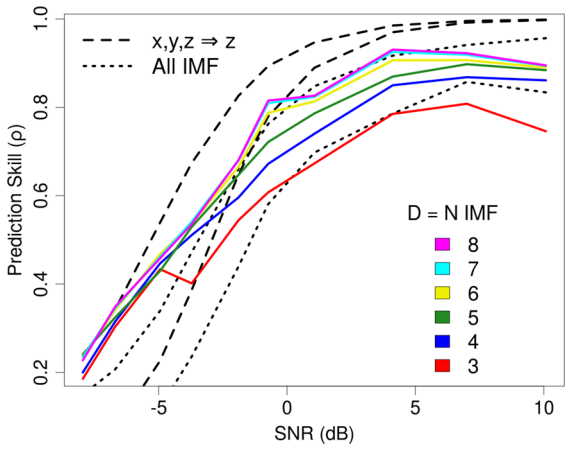

Fig. 4 Simplex projection skill of the most predictive IMFs of one noise realisation as a function of D, the number of IMFs. Dashed lines are the $\pm$ standard error envelopes of the 1000 noise ensembles of the reference model and state-spaces using all IMFs

and this is not a concern. Here, we use all available IMFs, and address IMF selection by modifying standard multiview embedding to use *out-of-sample* predictability as the metric for ranking top predictors.

The modified multiview embedding is performed using the complete set of IMFs from variables x and y with the training library of points [1-2000], prediction set [2001-3000], and forecast interval $T_P = 0$. The embedding dimension is set to $E = 1$ so that the IMFs are not time-delay embedded, but used explicitly as state-space variables. The most predictive combination of D IMFs are then selected to create a multivariable state-space.

Figure 4 plots simplex projection skill of the most predictive IMFs determined by multiview embedding for state-space dimensions of D from 3 to 8, comparing them to the $\pm$ standard error envelopes of the 1000 noise ensembles of the reference model and state-spaces using all IMFs. At SNR immediately below 3 dB the predictive skill of EMM with state-space dimension of $D = 6$ has converged to match the upper envelope of the full IMF state-space ensembles. At SNR below 0 (noise power exceeds signal power) the $D = 6$ projection exceeds that of the full IMF state-space.

To summarise, an $E = 3$ dimensional time-delay embedding on variables x and y to simplex project z from a state-space with dimension $D = 6$ matches that of the fully informed reference model ($[x, y, z]$ projecting z) at zero noise. As noise increases the Takens state-space model degrades in predictive skill, and,

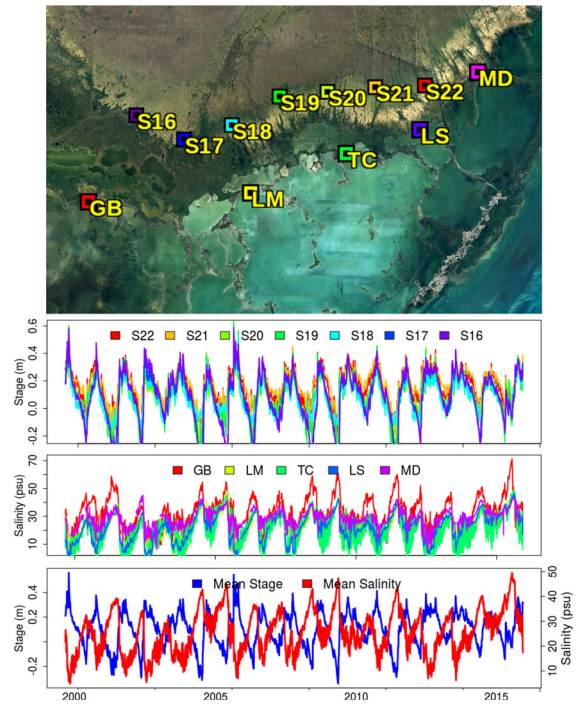

Fig. 5 Top: Aerial photograph of the southern Everglades and Florida Bay with hydrographic monitoring stations. Bottom: Station water levels (stage), salinity, and station-averaged water level and salinity over 16 years

both an indiscriminate use of all IMFs from noisy x and y, as well as a selected subset of $D = 6$ IMFs outperforms the Takens model at SNR less than 3 dB. We infer that EMM can provide improved state-space representations in the presence of noise.

4 Salinity in Florida Bay

Coastal South Florida is fringed by national parks including Biscayne and Everglades National Parks. Florida Bay is situated between the southern edge of the Florida peninsula and Florida Keys, is part of Everglades National Park, and, is an ecologically important and proliferent multispecies marine and estuarine nursery. Hydrologically, Florida Bay is interesting in that annual evaporation exceeds rainfall, resulting in widely varying coastal salinities as shown in Fig. 5. Hypersalinity is a seasonal occurrence, with episodic extremes surpassing 50 psu. These extreme hypersalinity events are associated with ecological collapse [24,25]. Clarification of influence variables and the ability to forecast salinities will directly inform ecosystem resource management.

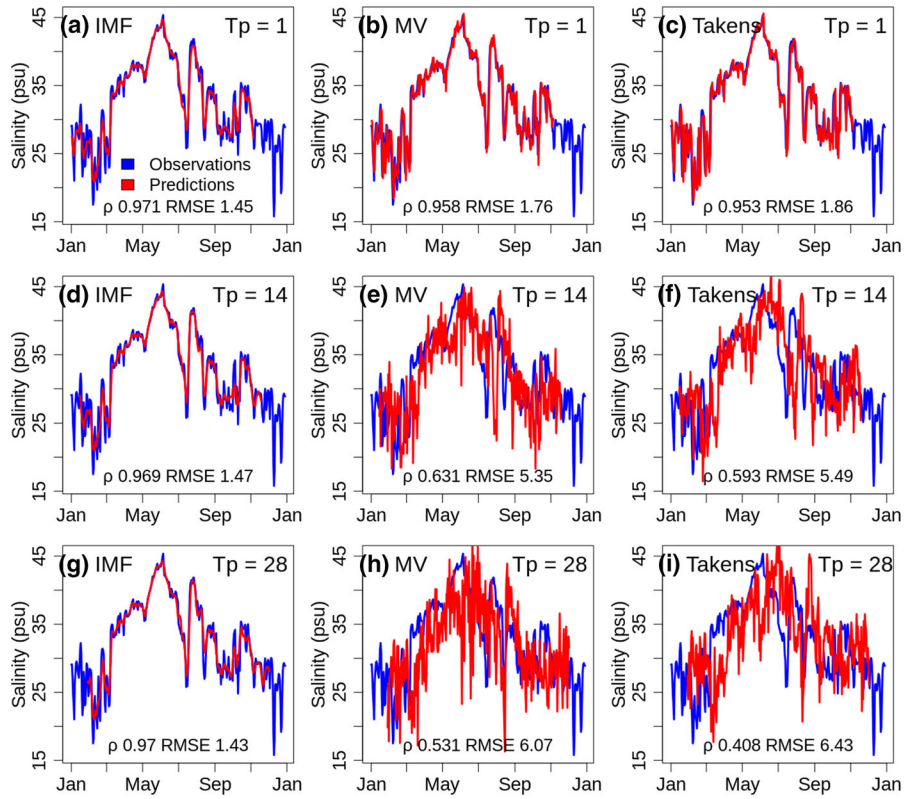

Fig. 6 Simplex in-sample projections of GB salinity for the year 2016 at different prediction intervals T_p for the EMM, multivariable, and Takens time-delay state-space representations

Our objective is to forecast salinity in Garfield Bight (GB), assessing predictive performance of EDM simplex using three different state-space representations described below. The Florida Bay data consist of 6332 days of mean water level (stage), salinity, evaporation, and rainfall. The lower panel of Fig. 5 illustrates a nonlinear inverse relationship between salinities and coastal marsh water levels signifying that water levels in the southern Everglades are a prime determinant of coastal salinities. Other causal variables are known to include evaporation, which is related to temperature and solar radiation, rainfall, water levels in the bay, and, previous salinity.

An initial step in EDM analysis is to examine the dimensionality and nonlinearity of the data [13], here, indicating that the data are nonlinear, and, a Takens time-delay embedding dimension of 5 provides good model predictability. Results of this data discovery are presented in "Appendix 1" (Sect. 1).

4.1 State-space models

The first state-space model is a Takens $E = 5$ time-delay embedding of GB salinity. Second, a multivariate state-space is constructed from the 5 variables GB salinity, stage, evaporation, rainfall, and, coastal land water level at station S16. These variables are known influencers of salinity as determined by the Florida Bay Assessment model [26]. The third state-space consists of selected IMFs of GB salinity, water level, evaporation, and coastal land water level at S16. IMFs are shown in "Appendix 1" (Sect. 1).

4.1.1 Model comparison

The comparative state-space analysis defines an in-sample data set as days [1-6330] for the training library, and days [5967-6275] as the prediction set. This prediction set corresponds to the period 2016-01-01 to 2016-11-04. The EMM model uses IMFs 5,6,7 of the salinity and stage variables, and IMFs 6,7,8 for S16 water level and evaporation (Fig. 12). These IMFs ignore high frequency, noise dominated components, as well interannual variations not deemed important on daily timescales.

Figure 6 plots a series of in-sample GB salinity projections for the three models over a sequence of pre-

diction horizons T_p. It is worth reinforcing that these are not predictions *per-se*, since the model is purely in-sample, rather, we seek to evaluate internal model representations of the dynamics. Here, we see that while the $T_p = 1$ projections are good for all three models, the EMM model misses high frequency variations evident in the January to March and September to October time frames. As prediction interval increases, these high frequency contributions seem to dominate the multivariate and time-delay embedding projections, significantly degrading accuracy, while the EMM model maintains a high fidelity internal representation of the dynamics.

4.2 Forecasting

We now turn to the question of whether EMM can be useful in an out-of-sample regime which is requisite for an applied forecast system. Simplex forecasts are made with the three state-space models from a training library consisting of points [1-5475] expanded in 30 day increments starting at day 5475 (2014-08-27) until the end of the record. Predictions are made at forecast intervals of 1 to 56 days in 7 day increments, starting at one day past the end of the training library. This 30-day moving window provides 28 samples at each prediction interval. The mean RMSE over each prediction interval is shown in Fig. 7 indicating that the EMM IMF state-space provides generally more accurate predictions of salinity than the multivariate and time-delay state-space models.

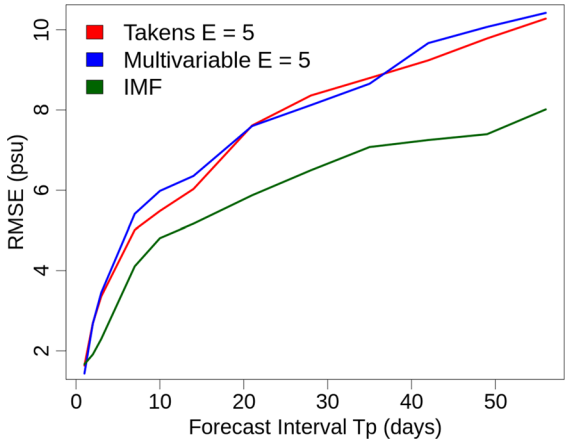

Fig. 7 Root mean square error (RMSE) of GB salinity out-of-sample simplex forecasts for the year 2016 at different prediction intervals T_p for the EMM, multivariable, and time-delay state-space representations

4.2.1 Comparison to numerical model

As noted earlier, salinity in coastal Florida Bay is an ecologically important factor, linked to the onset of widespread ecological collapse [24,25]. In 2015, environmental conditions aligned to produce historic record maximum salinities in excess of 70 psu (seawater is nominally 35 psu) followed by widespread sea-grass mortality and food-web disruption. The ability to forecast such events, or to attribute causal variables, provides actionable information for natural resource managers. Accordingly, equation-based physical models have been developed to explore these issues, and, the Bay Assessment Model (BAM) is one of the latest and best performing salinity models [26]. Here, we compare BAM and EMM forecasts for the 2015 hypersalinity event.

EMM forecasts are out-of-sample based on progressive training library and prediction horizon times. The training library starts with days [1-5721]. Predictions are made for 121 days starting at day 5722 (2015-5-1) over a range of prediction horizons T_p from 1 to 21 days. After each set of predictions, the training library end day and prediction start day are incremented by 1 day.

Figure 8 presents comparisons of the EMM forecasts with BAM results, where we note that BAM fails to capture the extreme hypersalinity event of 2015. EMM forecasts provide accurate salinity predictions at $T_p = 1$ with good forecasting of extreme salinity. Forecasts remain reasonably good at $T_p = 3$, having lost good fidelity at $T_p = 7$, but still outperforming the numerical model. This suggests that the method of progressive EMM projections at $T_p = 1$ are good candidates for an operational forecast system.

5 Conclusion

The 21st Century revolution in data collection and analysis has fundamentally changed our ability to infer relationships and dynamical behavior of complex systems. A powerful corollary is the emergence of data-driven, model-free analytical techniques where presumptions and models are not fit to observations, but where the data itself is represented in a state-space facilitating the discovery of dynamical relationships enabling state forecasting.

Inevitably, observational data are intertwined with noise, with potential to disrupt accurate state-space

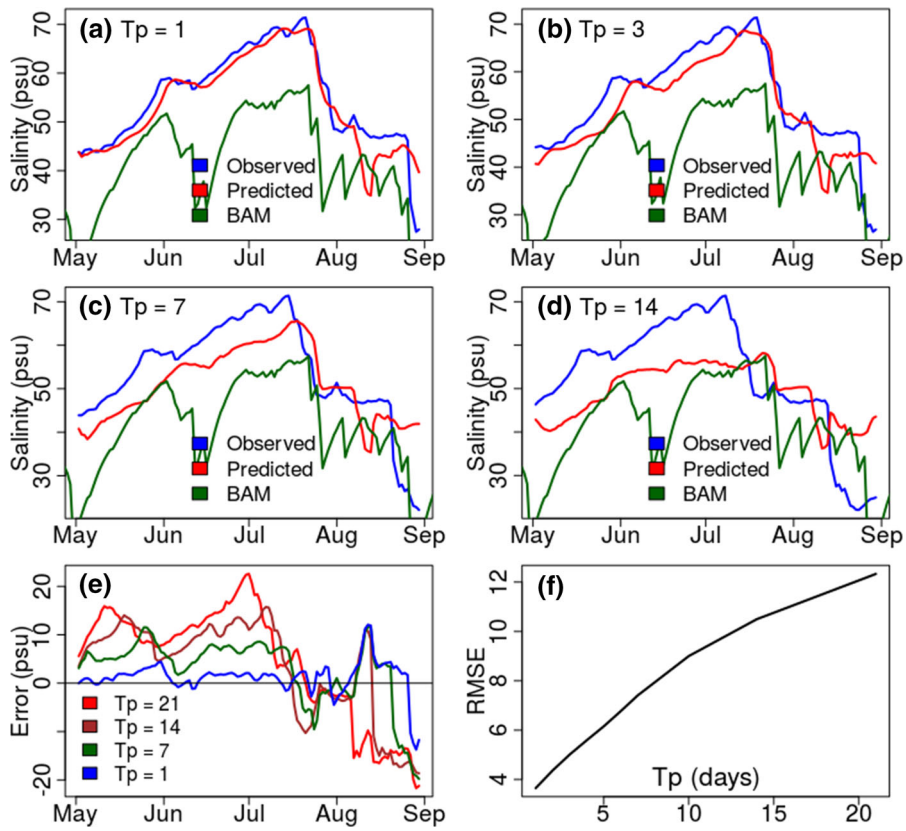

Fig. 8 Salinity forecasts in Garfield Bight during the 2015 hypersalinity event. **a**–**d** EMM forecasts (red) with observed data (blue) and results from the Bay Assessment Model (green). **e** EMM forecast error for different prediction horizons T_p. f) EMM RMS error over the prediction period May 1-August 31, 2015 as a function of forecast interval T_p

representations. Such disruptions motivate us to suggest empirical mode modeling (EMM), a synthesis of empirical mode decomposition (EMD) and empirical dynamic modeling (EDM). As the names suggest, these are data-driven, model-free techniques with potential for nonlinear dynamical analysis. EMM uses EMD intrinsic mode functions (IMF) as state-space vectors in the EDM framework. Since IMFs naturally decompose time series into physically relevant modes, they are well-suited for isolation and removal of noise.

EMM was assessed by application to a synthetic data set with controlled SNR, and, it was found that for high SNR state-spaces EMM as implemented here does not improve projection skill. However, as SNR falls below 3 dB EMM is found to improve model predictability and information content of the state-space, leading us to anticipate that noise-dominated signals can benefit from EMM.

Application to a geophysical data set composed of inherently noisy data found that EMM provided superior forecasting ability in comparison to time-delay or multivariate state-space representations. Further, the EMM representation outperformed the best available numerical model in predicting the 2015 hypersalinity event in Garfield Bight of Florida Bay, an extreme occurrence without precedent in the observational record.

Collectively, this work indicates that EMM can be a useful technique to improve state-space representations in the presence of noise with signal-to-noise ratios less than 3 dB. However, the 3 dB (half power) ratio may be data-specific, this is an area where further investigation is warranted.

Although this initial implementation and assessment of EMM seems promising, there are inherent limitations, and, avenues for improvement. The primary limitation is that EMD is not applicable to non-oscillatory signals. Time series such as neuronal spike trains, rainfall, or other discrete, Poisson process dynamics will not be amenable to EMM based on EMD IMFs. This might be addressed by replacing the EMD IMF with a modal decomposition that captures non-oscillatory time-dependence, for example, a discrete wavelet transform.

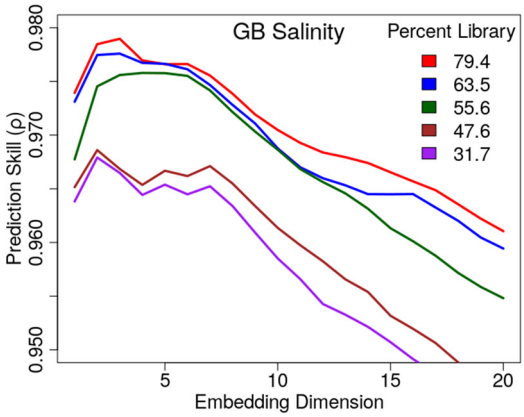

Fig. 9 Out-of-sample simplex model correlation for Florida Bay station GB salinity as a function of time-delay embedding dimension. Different curves are based on different library sizes, expressed as a percentage of the total data size. The maximum library index is 5000, predictions are over days 5001-6300

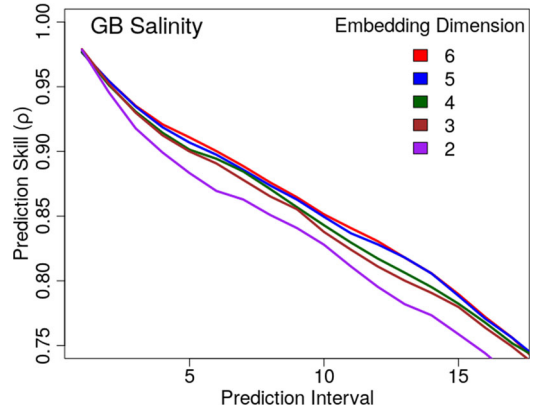

Fig. 10 Out-of-sample simplex model correlation for Florida Bay station GB salinity as a function of forecast interval of T_p

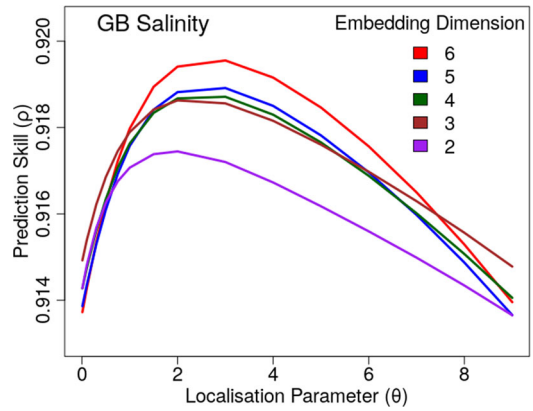

Fig. 11 Out-of-sample s-map model correlation for Florida Bay station GB salinity as a function of nonlinear localization parameter θ

As mentioned earlier, a presumption of EMM is that dynamically-relevant IMF's have been selected from the EMD. In cases where the IMF selection excludes system dynamics, there is potential to miss dynamical features of interest. This is an on-going area of investigation.

Regarding improvements, investigating methods to optimally select IMFs to avoid noise, or, to maximize cross-variable information assessments are warranted. Here, we have relied on EDM multiview embedding. Second, EMD has been extended beyond the original decomposition algorithm used here. For example, ensemble empirical mode decomposition (EEMD) has been found to mitigate mode-mixing and improve physical significance of modes [27]. EMD has also been extended into hybrid machine learning paradigms [28,29] and multivariate implementations [30].

Another obvious exploration is to assess complementary mixed embeddings as a state-space representation. For example, combinations of Takens time-delay embedded vectors with IMFs, and, with multivariate observations. Yet another is to use other nonlinear modal decompositions such as the discrete wavelet transform either in-place of, or, as a complement to IMFs. Finally, the use of sequential locally weighted global linear maps (S-maps) [7] may provide improved state-space representation and forecast skill.

Funding This work was funded in collaboration of the U.S. Department of the Interior and University of California San Diego through the Cooperative Ecosystem Studies Units (CESU) Network, and, the Provisional Technical Secretariat of the Comprehensive Nuclear-Test-Ban Treaty Organization. This work was also supported by the McQuown Fund, and, the McQuown Chair in Natural Sciences, University of California, San Diego. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability Statement The datasets generated during and analysed during the current study are available from the corresponding author on reasonable request.

Conflict of interest The authors declare that they have no conflict of interest. The views expressed herein are those of the authors and do not necessarily reflect the views of the CTBTO Preparatory Commission.

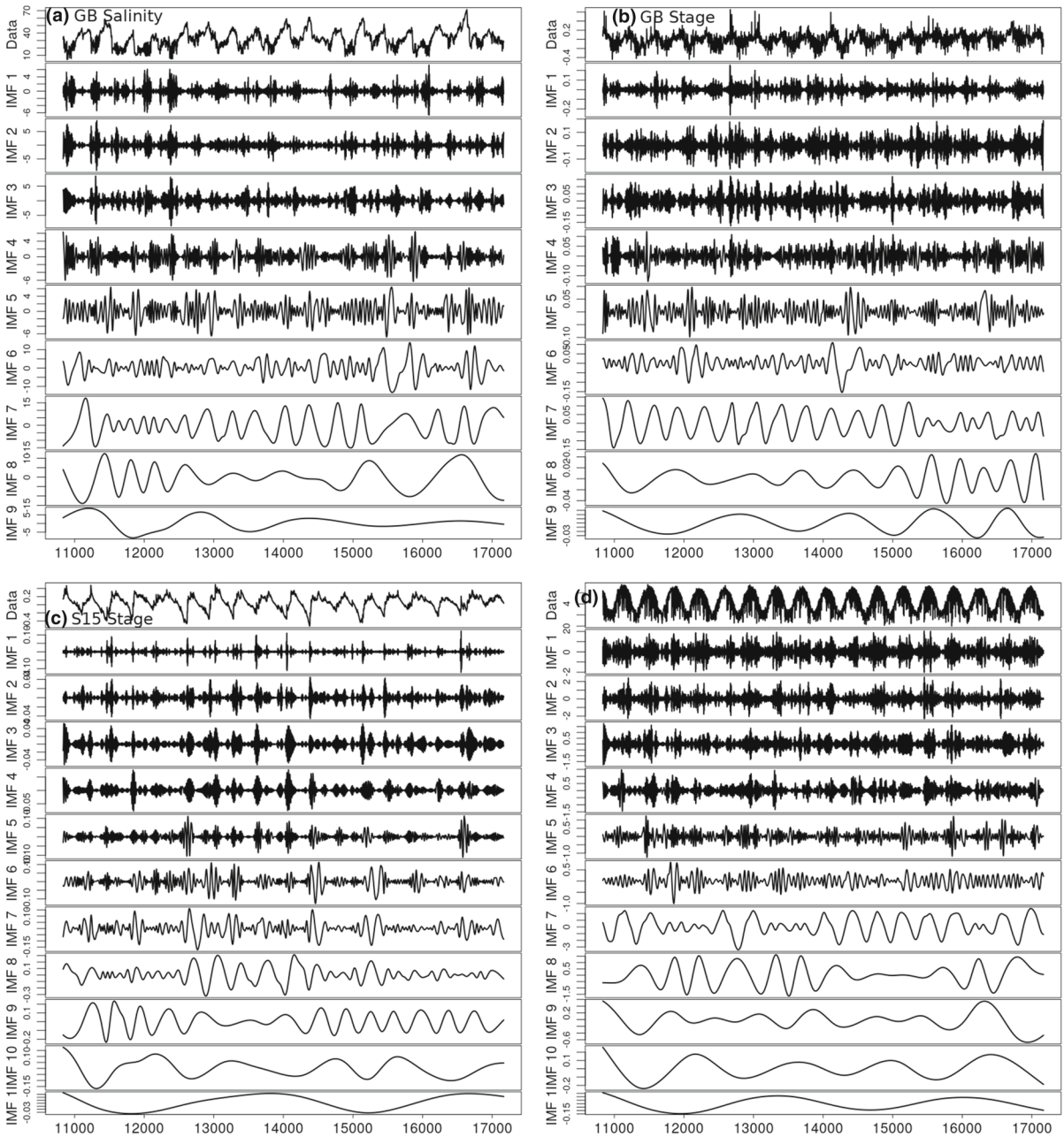

Fig. 12 EMD IMFs of variables influencing Garfield Bight salinity. **a** Salinity. **b** Water level. **c** Coastal land water level. **d** Evaporation

Appendix 1: Florida Bay salinity data discovery

Application of EDM to a dataset is predicated on assessing whether the underlying dynamics are state-

dependent (nonlinear), and if so, estimating an embedding dimension to use if univariate data are to be time-delay embedded. A primer on EDM analysis can be found in [13].

The Florida Bay data consist of 6332 daily values (September 1 1999 though December 31 2016). We use days 1-5000 as the training (library) set, and days 5001–6300 as the prediction.

Figure 9 shows simplex prediction skill as a function of Takens embedding dimension for Garfield Bight (GB) salinity. Predictions are made at a forecast interval of $T_p = 1$ timestep ahead. This suggests that embedding dimensions in the range of 2-6 provide the best model fidelity, and that the underlying dynamics of the salinity are low-dimensional.

Figure 10 presents simplex prediction skill as a function of forecast interval T_p at different embedding dimensions E. The saturation of predictability as E approaches 5 indicates that a dimension of 5 can be a reasonable choice for EDM analysis of this data. The decline in predictability as T_p increases is consistent with dynamics of a nonlinear system.

Interpreting nonlinearity as state-dependence [7], the EDM s-map algorithm allows one to assess nonlinearity with examination of predictive skill as a function of state-space linear localization (θ). Figure 11 shows this assessment for the GB salinity data, indicating a weak, but robust state-dependence, again indicative that the salinity dynamics are nonlinear.

Figure 12 presents EMD IMFs of physical variables influencing salinity in Garfield Bight, Florida Bay. Low order, high frequency IMFs are considered noise dominated and not used as EDM state-space vectors. High order, low frequency IMFs are also not used as state-space variables.

References

1. Anderson, P.W.: More is different. Science **177**(4047), 393–396 (1972). https://doi.org/10.1126/science.177.4047.393
2. See for example: *Chaos: An Interdisciplinary Journal of Nonlinear Science*, American Institute of Physics (AIP), ISSN 1054-1500, 1089-7682. https://aip.scitation.org/journal/cha; *Complex Systems*, Complex Systems Publications, Inc., ISSN 0891-2513. https://www.complex-systems.com/; *Physical Review E*, American Physical Society, ISSN 2470-0045, 2470-0053 https://journals.aps.org/pre/
3. DeAngelis, D.L., Yurek, S.: Equation-free modeling unravels the behavior of complex ecological systems. PNAS **112**(13), 3856–3857 (2015). https://doi.org/10.1073/pnas.1503154112
4. Lin, B., He, X., Ye, J.: A geometric viewpoint of manifold learning. Appl. Inf. **2**, 3 (2015). https://doi.org/10.1186/s40535-015-0006-6
5. Coifman, R., Stéphane, L.: Diffusion maps. Appl. Comput. Harmon. Anal. **21**(1), 5–30 (2006). https://doi.org/10.1016/j.acha.2006.04.006
6. Sugihara, G., May, R.: Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature **344**, 734–741 (1990). https://doi.org/10.1038/344734a0
7. Sugihara, G.: Nonlinear forecasting for the classification of natural time series. Philos. Trans. Phys. Sci. Eng. **348**(1688), 477–495 (1994)
8. Dixon, P.A., Milicich, M., Sugihara, G.: Episodic fluctuations in larval supply. Science **283**, 1528–1530 (1999). https://doi.org/10.1126/science.283.5407.1528
9. Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E., Fogarty, M., Munch, S.: Detecting causality in complex ecosystems. Science **338**, 496–500 (2012)
10. Ye, H., Sugihara, G.: Information leverage in interconnected ecosystems: overcoming the curse of dimensionality. Science **353**, 922–925 (2016)
11. Takens, F.: Detecting strange attractors in turbulence. In: Rand, D.A., Young, L.-S. (eds.) Dynamical Systems and Turbulence. Lecture Notes in Mathematics, vol. 898, pp. 366–381. Springer-Verlag, Berlin (1981)
12. Deyle, E., Sugihara, G.: Generalized theorems for nonlinear state space reconstruction. PLoS ONE **6**(3), e18295 (2011). https://doi.org/10.1371/journal.pone.0018295
13. Chang, C., Ushio, M., Hsieh, C.: Empirical dynamic modeling for beginners. Ecol. Res. **32**, 785–796 (2017). https://doi.org/10.1007/s11284-017-1469-9
14. Huang, N.E., Wu, Z.H.: A review on Hilbert-Huang transform: method and its applications to geophysical studies. Rev. Geophys. **46**(2), RG2006 (2008). https://doi.org/10.1029/2007RG000228
15. Dai, W., Tang, L., Yu, L.: Why do EMD-based methods improve prediction? A multiscale complexity perspective. J. Forecast. **38**(7), 714–731 (2019). https://doi.org/10.1002/for.2593
16. Looney, D., Hemakom, A., Mandic, D.: Intrinsic multiscale analysis: a multi-variate empirical mode decomposition framework. Proc. R. Soc. A. **471**, 2173 (2015). https://doi.org/10.1098/rspa.2014.0709
17. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica **37**(3), 424–438 (1969). https://doi.org/10.2307/1912791
18. Molkov, Y., Loskutov, B., Mukhin, D., Feigin, A.: Random dynamical models from time series. Phys. Rev. E **85**, 036216 (2012). https://doi.org/10.1103/PhysRevE.85.036216
19. Munch, S., Brias, A., Sugihara, G., Rogers, T.: Frequently asked questions about nonlinear dynamics and empirical dynamic modelling. ICES J. Mar. Sci. **77**(4), 1463–1479 (2020). https://doi.org/10.1093/icesjms/fsz209
20. Arnold, L.: *Random Dynamical Systems*. Springer Monographs in Mathematics, p. 586. Springer-Verlag, Berlin Heidelberg. (1998) http://doi.org/10.1007/978-3-662-12878-7

21. Gavrilov, A., Loskutov, E., Mukhin, D.: Bayesian optimization of empirical model with state-dependent stochastic forcing. Chaos Solitons Fract. **104**, 327–337 (2017). https://doi.org/10.1016/j.chaos.2017.08.032
22. Rössler, O.E.: An equation for continuous chaos. Phys. Lett. **57A**(5), 397–398 (1976). https://doi.org/10.1016/0375-9601(76)90101-8
23. Timmer, J., König, M.: On generating power law noise. Astron. Astrophys. **300**, 707–710 (1995)
24. Hall, M., Furman, B., Merello, M., Durako, M.: Recurrence of Thalassia testudinum seagrass die-off in Florida Bay, USA: initial observations. Mar. Ecol. Prog. Ser. **560**, 243–249 (2016). https://doi.org/10.3354/meps11923
25. Johnson, C.R., Koch, M.S., Pedersen, O., Madden, C.J.: Hypersalinity as a trigger of seagrass (Thalassia testudinum) die-off events in Florida Bay: evidence based on shoot meristem O2 and H2S dynamics. J. Exp. Mar. Biol. Ecol. **504**, 47–52 (2018). https://doi.org/10.1016/j.jembe.2018.03.007
26. Park, J., Stabenau, E., Kotun K.: "Florida Bay Assessment Model: User Manual". South Florida Natural Resources Center, U.S. Department of the Interior, Everglades National Park, Homestead, FL. Hydrologic Model Manual. SFNRC 2016:7-27. 62 pp. (2016) https://github.com/SoftwareLiteracyFoundation/BAM
27. Wu, Z., Huang, N.: Ensemble empirical mode decomposition: a noise-assissted data analysis method. Adv. Adapt. Data Anal. **01**(01), 1–41 (2009). https://doi.org/10.1142/S1793536909000047
28. Chen, Q., Wen, D., Li, X., Chen, D., Lv, H., Zhang, J., et al.: Empirical mode decomposition based long short-term memory neural network forecasting model for the short-term metro passenger flow. PLoS ONE **14**(9), e0222365 (2019). https://doi.org/10.1371/journal.pone.0222365
29. Jiang, C., Conde, M., Deng, B., Chen, J.: Hybrid improved empirical mode decomposition and BP neural network model for the prediction of sea surface temperature. Ocean Sci. **15**, 349–360 (2019). https://doi.org/10.5194/os-15-349-2019
30. Rehman N., Mandic D. P.: "Multivariate empirical mode decomposition". In*Proceedings of The Royal Society of London A: Mathematical, Physical and Engineering Sciences*, pg. rspa20090502 (2009)
31. Aimon, S., Katsuki, T., Jia, T., Grosenick, L., Broxton, M., Deisseroth, K., et al.: Fast near-whole-brain imaging in adult Drosophila during responses to stimuli and behavior. PLoS Biol. **17**(2), e2006732 (2019). https://doi.org/10.1371/journal.pbio.2006732
32. Jutten, C., Hérault, J.: Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture. Signal Process. **24**, 1–10 (1991)
33. Ahrens, M.B., Orger, M.B., Robson, D.N., Li, J.M., Keller, P.J.: Whole-brain functional imaging at cellular resolution using light-sheet microscopy. Nat. Methods **10**, 413–420 (2013)