Environmental Modelling & Software 81 (2016) 1–11

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

System learning approach to assess sustainability and forecast trends in regional dynamics: The San Luis Basin study, Colorado, U.S.A

Alejandra M. González-Mejía [1], Tarsha N. Eason, Heriberto Cabezas*

Sustainable Technology Division, National Risk Management Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, 26 W. Martin Luther King Drive, Cincinnati, OH 45268, USA

ARTICLE INFO

Article history:
Received 18 May 2015
Received in revised form
31 October 2015
Accepted 2 March 2016
Available online 22 March 2016

Keywords:
Artificial neural network
Fisher information
Forecast
Prediction
Baseline scenario
Sustainability
Regional system

ABSTRACT

This paper presents a methodology that combines the power of an Artificial Neural Network and Information Theory to forecast variables describing the condition of a regional system. The novelty and strength of this approach is in the application of Fisher information, a key method in Information Theory, to preserve trends in the historical data and prevent over fitting projections. The methodology was applied to demographic, environmental, food and energy consumption, and agricultural production in the San Luis Basin regional system in Colorado, U.S.A. These variables are important for tracking conditions in human and natural systems. However, available data are often so far out of date that they limit the ability to manage these systems. Results indicate that the approaches developed provide viable tools for forecasting outcomes with the aim of assisting management toward sustainable trends. This methodology is also applicable for modeling different scenarios in other dynamic systems.

Published by Elsevier Ltd.

1. Introduction

Determining how to assess and manage aspects of a system towards a sustainable path is one of the most critical challenges in sustainability science (Kates, 2011). Accordingly, developing future scenarios is an important management need (Boyko et al., 2012). A key element in this effort is developing plausible projections given an assessment of historical data to identify trends representing typical system conditions. Artificial neural networks offer great promise for handling time series forecasting and form the basis of our approach.

Artificial neural networks (ANN) are powerful data-driven modeling techniques based on iterative algorithms that have the ability to estimate a function from a great array of dependent or independent inputs (Zaihong et al., 2012). They provide an advantage over many traditional statistical approaches because there is no need to find a causal relationships among variables. ANNs are useful in a variety of applications from data processing and regression analysis to adaptive learning and forecasting, yet like any estimation technique, ANNs are prone to over-fit forecasted data (Krogh, 2008). In response to this challenge, we developed a methodology to “bound” the forecast and ensure that ANN projections describe typical patterns found in the historical data. Here, we use Fisher information to characterize system condition as defined by patterns in variables.

Fisher information is a measure of the amount of information about system behavior that is present in data (Fisher, 1922). While, its roots are in statistical estimation theory, Fisher information has been adapted to provide a means of monitoring variables which characterize system behavior by collapsing them into an index that can be monitored to assess dynamic order (patterns in data), system regimes and regime shifts (Fath et al., 2003; Karunanithi et al., 2008). For this application, Fisher information was initially employed to assess the stability of the patterns (how much they are changing) in the historical data. The trends in the index also served as a constraint to ensure that level of stability found in the historical data is preserved during the time series projection. The central purpose of this research effort is to develop a data-driven forecasting method that ensures that the projections contain patterns

* Corresponding author.
E-mail addresses: gonzalez.alejandra@epa.gov (A.M. González-Mejía), eason.tarsha@epa.gov (T.N. Eason), cabezas.heriberto@epa.gov (H. Cabezas).

[1] Research Fellow, Oak Ridge Institute for Science and Education and Research Fellow presently at Bangor University, Wales, UK. Sêr Cymru National Research Network for Low Carbon, Energy and Environment, School of Environment, Natural Resources and Geography.

http://dx.doi.org/10.1016/j.envsoft.2016.03.002
1364-8152/Published by Elsevier Ltd.

consistent with the historical behavior of the system.

As a study case, we applied the method to data collected on a regional system in south-central Colorado. This work is a follow up to the San Luis Basin (SLB) Sustainability Metrics Project initiated in 2006 by the U.S. Environmental Protection Agency (USEPA), Office of Research and Development (USEPA, 2010). Using publically available data and four integrated metrics, researchers computed these metrics to assess trends in key aspects of sustainability: ecological footprint representing environmental burden (Hopton and White, 2012), green net regional product accounting for economic well-being (Heberling et al., 2012), emergy which measures energy resources and flow through the system (Campbell and Garmestani, 2012), and Fisher information captures the stability of the system condition (Eason and Cabezas, 2012). The research team found that the dynamics of the region were fairly steady yet slowly moving away from sustainability. The stability was due to such drivers as relatively low diversity in land use, high bioavailability and a relatively stable economic base (Hopton et al., 2010). One of the primary goals of the study was help inform decision making on land and environmental management issues. However, data availability presented a key limitation and resulted in the production of sustainability metrics with a lag of three or more years. Therefore, these calculations were not the optimal for stakeholders to make near-time informed management decisions (Heberling and Hopton, 2014). Accordingly, the team highlighted the need for forecasting methodologies to aid in assessing plausible scenarios and alternative futures. This research article presents a method of developing and applying a forecasting methodology to aid in producing scenarios for the SLB. The approach presented teams a dynamic autoregressive artificial neural network (DARX) with Fisher information in order to extrapolate time series from 2011 to 2025 while preserving the trends found from 1969 to 2010. This approach simulates a "business as usual" scenario that eliminates both unexpected patterns and model over fitting. Although this initial installment focuses on developing a baseline scenario, the methods can be adapted to assess alternate futures and other types of systems.

2. Methods

The core method involves two main processes: (1) Characterizing the system given data series from 1969 to 2010 and (2) forecasting of a baseline scenario for the period 2011–2025. Process 1 involves using Fisher information to assess trends in the system variables. Process 2 consists in designing (simulation, calibration, cross-validation, test, and optimization) and applying the architecture of the constrained neural net to extrapolate trends found in process 1. Both processes incorporate techniques to handle data quality and quantity issues (e.g., sparseness) inherent in real systems by pre and post processing the variables describing an area of study.

2.1. Artificial neural nets

Inspired by biological neural networks, researchers from different disciplines design artificial neural nets in order to address a variety of problems, such as pattern classification, clustering, optimization, and prediction among others (Jain et al., 1996). The analogy between the artificial and the biological system is the high capacity of interconnection (neurons), learning (what is more probable), generalizing (rules), and making decisions (model). These characteristics provide a powerful tool for answering questions about the future, based on past behavior. In that sense, neural nets are an excellent forecasting choice that does not need prior knowledge of the relationship among data (not always evident in observations), and that infers from examples in spite of noise content (Darbellay and Slama, 2000). Further, neural nets have been applied successfully to model a wide variety of real-world applications, such as: the prediction of the stock market index (Guresen et al., 2011), forecasting energy consumption (Kankal et al., 2011) and estimating wind power output (Tu et al., 2012). Other examples include the forecast of environmental variables such as the net ecosystem metabolism (a proxy for system tropic state) within a freshwater wetland (Young II et al., 2011), water quality (Zaihong et al., 2012), and municipal waste generation (Antanasijević et al., 2013). As mentioned, Artificial Neural Networks (generally called neural nets) have demonstrated promise in the forecasting arena (Maier et al., 2010; Guresen et al., 2011; Zaihong et al., 2012; Antanasijević et al., 2013). However, when used for projections, they are sometimes prone to over fitting and produce results that are outside of the realm of realistic system behavior (Voyant et al., 2011). Hence, it is critical that mechanisms be developed to help establish appropriate constraints on projections.

2.2. Fisher Information

Integrated indicators or metrics have proven a very useful tool to present a scientific, straightforward, consistent, and multidimensional view of a system (Bond and Morrison-Saunders, 2010). One critically important characteristic is the patterns in variables describing the condition of the system. Fisher information (Fisher, 1922), denoted I in Equation (1), was selected to measure the dynamics and stability (Pawlowski et al., 2005) of the system under study.

$$I = \int \frac{1}{p(s)} \left[\frac{dp(s)}{ds} \right]^2 \mathrm{d}s \tag{1}$$

Derived from Information Theory, Fisher information is a statistical quantity that measures the degree to which a parameter (e.g., s: state of a system) can be estimated (Frieden, 2004; Mayer et al., 2007) from a given data set, where p(s) is the probability of observing a particular state (condition) of the system. Fisher information (I) has been adapted to assess dynamic order (patterns in data) and applied to various types of systems for evaluating such aspects as urban and regional sustainability (Eason and Cabezas, 2012; Gonzalez-Mejia et al., 2012; Gonzalez Mejia et al., 2014). Further details on the derivation and computation of the index may be found in Karunanithi et al., 2008 and the USEPA (2010). Simply stated, higher I implies greater predictability of the system's state at a point in time (Mayer, 2008) and higher information content in the data set.

According to the Sustainable Regimes Hypothesis, well-functioning systems are predictable. The patterns may fluctuate within a limited range exist in regimes but they remain relatively steady through time; accordingly, the time averaged Fisher Information is constant, ($dI/dt \approx 0$) (Karunanithi et al., 2008). Taking this a bit further, a system is considered to be in a stable regime when values are within two standard deviations ($2sd(I)$) from the mean value I computed for the historical data (Equation (2)). A regime shift exists only when the I drops by more than two standard deviations from I as shown in Equation (3).

$$\text{Stable period} \Rightarrow \langle I \rangle - 2 \cdot sd(I) \leq I(t) \leq I + 2 \cdot sd(I) \tag{2}$$

$$\text{Regime shift} \Rightarrow I(t) < I - 2 \cdot sd(I) \tag{3}$$

The choice of two standard deviations from the mean as a criteria is based on an application of a Chebyshev's Theorem, which states that the proportion of observations falling within k standard deviations from the mean is at least $1-1/k^2$ for any arbitrary probability distribution (Lapin, 1975). Thus, at least 75% of the I values are within two standard deviations from the mean (i.e., typical behavior) and by exclusion, outliers can be identified.

2.3. Approach

The approach for examining the SLB system includes: 1) using I to characterize the patterns in the system (Fig. 2a) and 2) designing and applying neural nets to forecast the system variables (Fig. 2b). The approach presented in this work is unique in its use of a sustainability index based on information theory for bounding the projection such that the extrapolation is in line with the patterns present in the historical data. Further, it ensures that information content in the system is preserved. This methodology allows us to establish a base case study of the San Luis Basin Colorado region useful to estimate scenarios.

2.3.1. System characterization

The San Luis Basin (SLB) is a primarily rural, agricultural region with a low population density (Table 1). This seven county region contains the Upper Rio Grande River Basin, the San Luis Valley, and the Great Sand Dunes National Park and Preserve (Fig. 1). The area has distinct hydrological boundaries, a large amount of publicly owned land (approximately 78%) local government support and strong community interest in the sustainable development of the region (Hopton et al., 2010). In line with this pilot study, we selected and collected time series data that describe the demographic, economic, consumption (energy and food) production and environmental components of the system (Table 2 and Table A1). Due to data availability issues, several strategies were employed to gather the data for the region. Table 2 presents a summary of the twenty seven variables used in the assessment. Details of the data collection are provided in Appendix A.

Characterizing the system consisted of using Fisher information to assess the dynamics of each system component (Demographic (D), environmental (E), energy consumption (EC), food consumption (FC1 and FC2), and production (P), as well as for the overall system including all time series in Table 2 from 1969 to 2010 to capture the inherent information in past trends of the input variables (Fig. 2a). Then, stable periods (Equation (2) and regime shifts (Equation (3) defining sustainable trends in the SLB were assessed for each component and the overall system.

2.3.1.1. Preprocessing of data and fisher information calculation. In order to make the process of learning more efficient, it is recommended that neural net inputs are normalized (Beale et al., 2013). Preferably, these time series should contain many time points for training (Guresen et al., 2011). For that reason, each of the 27 variables was divided by its maximum value in the series (1969–2010) and interpolated quarterly with a piecewise cubic hermite polynomial function (PCHIP, Matlab®) to get 165 normalized data points per variable.

Since, data are preprocessed, it is also imperative to assure that the information content of the original variables is the same as the normalized and interpolated data set. Thus, this index was computed for 27 original variables and also for the 27 preprocessed (normalized and interpolated) variables which will be used to extrapolate the components of the SLB regional system. Results of two tests (i.e., linear correlation adjusted $R^2 = 0.8$, and nonparametric regression spearman rank order correlation

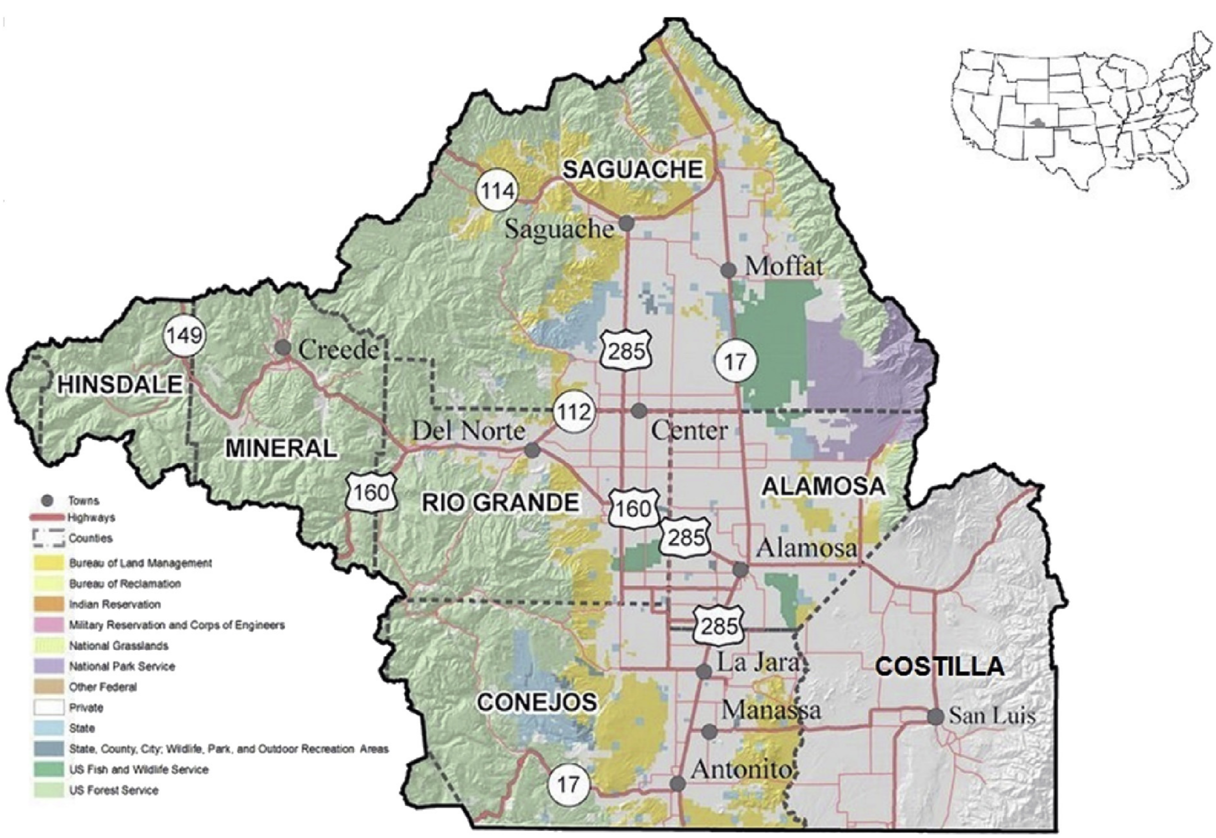

Fig. 1. Area of study. Slight modification of map (Bureau of Land Management-Colorado/San Luis Valley Field Office, 2014).

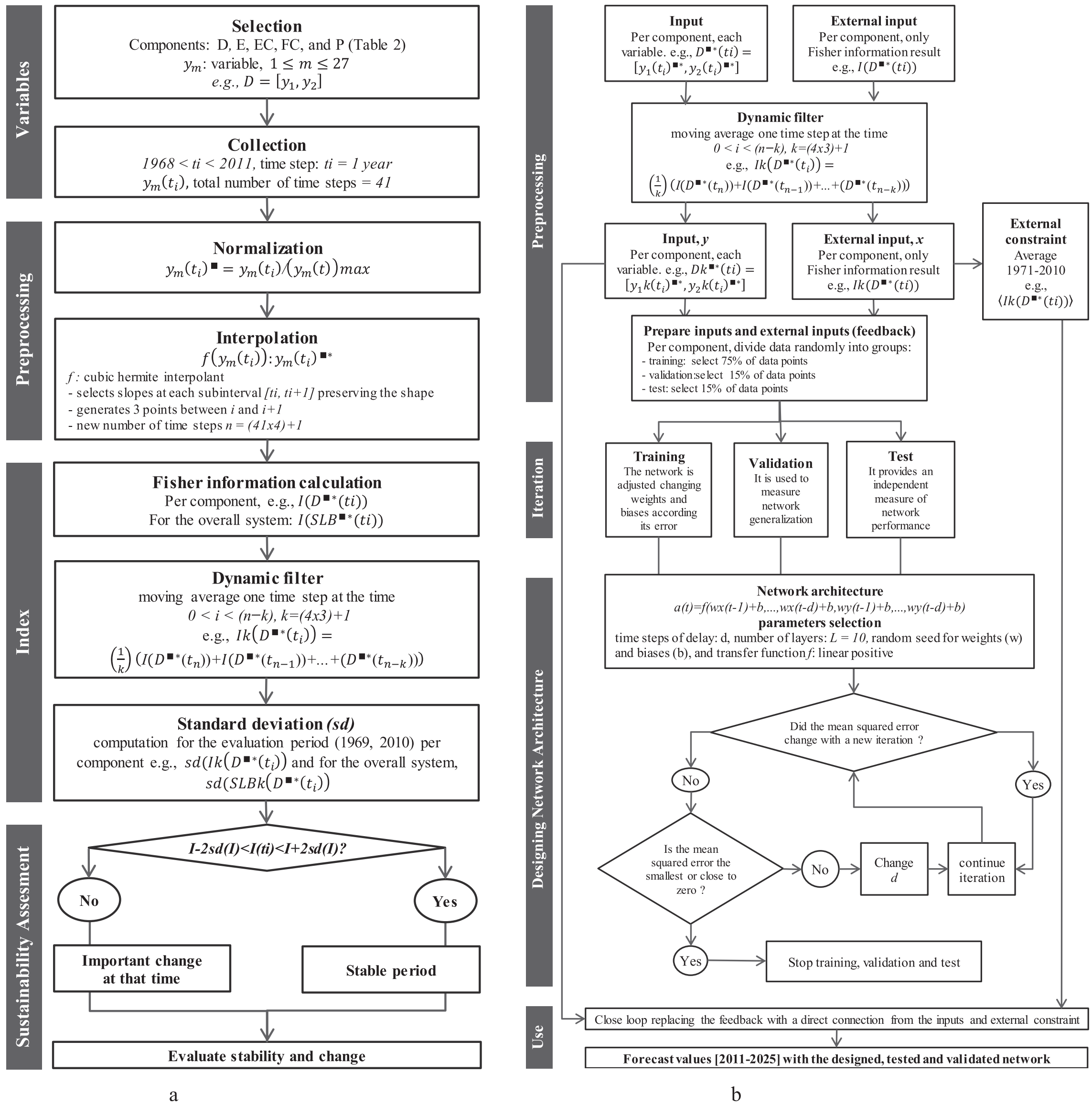

Fig. 2. a) System characterization. In order to characterize the system (Fig. 2a), the first step is the selection and collection of variables. For this step, consider the five components on Table 2, each one has $y_m(t_i)$ variables with $1 \le m \le 27$ and $1969 \le t_i \le 2010$: Demographic (D), $1 \le m \le 2$, environmental (E), $3 \le m \le 5$, energy consumption (EC), $6 \le m \le 10$ food consumption (FC), $11 \le m \le 22$ and production (P), $23 \le m \le 27$. Each time series is normalized by its maximum value, $(y_m(t))max$, and then interpolated using a cubic hermite polynomial (PCHIP®) to get 165 data points in total for a preprocessed variable, $y_m(t_i)^{■*}$. Then, Fisher information is calculated per component, e.g., $y_1(t_i)^{■*}$ and $y_2(t_i)^{■*}$ were used to compute $I(D^{■*}(ti))$, and for the overall system that includes the 27 variables, $I(SLB^{■*}(ti))$. Note that the square indicates normalization and the asterisk denotes interpolation of the respective times series. After the I calculation, all the variables were processed again with a dynamic filter of 12 data points (k=12), moving the window one time step at the time, e.g., $Ik(D^{■*}(t_i))$. Lastly, the standard deviation of each component is calculated to define stable periods and if present, regimes shifts (Equations (2) and (3)) in order to assess sustainable trends in the SLB. b. Forecast process. Preprocessing of data and Fisher information.

Table 1
San Luis Basin, Colorado, 2010 population.

U.S. Census, 2010	State	Region	Counties in the San Luis Basin region						
	Colorado	SLB	Alamosa	Conejos	Costilla	Hinsdale	Mineral	Rio Grande	Saguache
Population	5,029,196	46,870	15,445	8256	3524	843	712	11,982	6108
Land area (square miles)	103,642	9310	723	1287	1227	1117	876	912	3169
Persons per square mile	49	5	21	6	3	1	1	13	2

(U.S. Census Bureau, 2013).

coefficient = 0.9) corroborate that the trends in both data sets are strongly correlated therefore, the information content is essentially the same (Appendix C). Appendix C also shows that the system was in a stable regime from 1969 to 2010 according to the criterion in Equation (2). This process ensures that the data manipulation has not created information (i.e., preserves the trends in the original set) and confirms that the preprocessed data maintains the original data.

Table 2
Components and variables of the system. Appendix A.1 contains the annual values for each variable describing the agricultural region in the SLB, Colorado and Appendix A.2 consolidates the environmental stations in the area and period of study.

Component		Variable			Availability		Source	Data link	Notes
		#	Name	Unit	Level	Years			
DEMOGRAPHIC (D)		1	Population (P)	Inhabitants (Inhab./year)	County	1969–2011	(U.S. Bureau of Economic Analysis, 2013)	http://www.bea.gov/i.Table/index_regional.cfm	SLB: Counties addition (Alamosa + Conejos + Costilla + Mineral + Hinsdale + Rio Grande + Saguache)
		2	Personal income (PI)	Thousands of dollars (KUSD$/year)		1969–2011			
ENVIRONMENTAL (E)		3	Precipitation (R)	Inches (in/year)	Station	1948–2011	National Oceanic and Atmospheric Administration NOAA (National Climatic Data Center, 2013)	http://www.ncdc.noaa.gov/data-access/quick-links	SLB: monthly mean of all stations (Table A.2) added for precipitation and averaged for temperature (Jan–Dec)
		4	Temperature (T)	Fahrenheit (F/year)		1948–2011			
		5	Carbon dioxide emissions (CO2)	Million Metric Tons (MMT/year)	National	1969–1979	(U.S. Energy Information Administration, 2013c)	http://www.eia.gov/environment/emissions/state/state_emissions.cfm	SLB $_{1969-1979}$ = [$SLBEC_{FF}/USEC_{FF}$] × $USCO_2$
					State	1980–2010	(U.S. Energy Information Administration, 2013a)		SLB $_{1980-2010}$ = [$SLBEC_{FF}$/$ColoradoEC_{FF}$] × $ColoradoCO_2$
CONSUMPTION	ENERGY (EC)	6	Coal	Trillion Btu (TBTU/year)	State	1960–2011	(U.S. Energy Information Administration, 2013b)	http://www.eia.gov/state/seds/seds-data-complete.cfm?sid=CO	SLBEC = Colorado energy consumption per capita (ColoradoEC/Inhab.) × SLBP $EC_{FF} = EC_{coal} + EC_{natural\ gas} + EC_{petroleum}$
		7	Natural Gas			1960–2011			
		8	Petroleum			1960–2011			
		9	Hydro-electric			1960–2011			
		10	Wood and waste			1960–2011			
	FOOD Part 1: variables 11–16 (FC1) Part 2: variables 17–22 (FC2)	11	Fruit and vegetables	Pounds per year (lb/year)	Country	1970–2010	United States Department of Agriculture (USDA) (Economic Research Service, 2013)	http://www.ers.usda.gov/data-products/food-availability-(per-capita)-data-system.aspx	Total fruits and vegetables (fresh weight equivalent). 1969 assumed value = (1970 + 1971)/2
		12	Dairy			1909–2010			All dairy products (milk equivalent)
		13	Flour and cereal			1909–2010			Total flour and cereal products includes wheat and rye flour, rice, corn, oat and barley products
		14	Caloric sweeteners			1909–2010			Total caloric sweeteners (dry weight) includes refined cane and beet sugar, corn sweeteners, edible syrups, and honey
		15	Potatoes			1909–2010			All potatoes (farm)
		16	Red meat			1909–2010			Total red meat (boneless weight) includes beef, veal, pork, and lamb
		17	Added fats and oils			1909–2010			Total added fats and oils (product weight)
		18	Poultry			1909–2010			Total poultry (boneless weight) includes chicken and turkey
		19	Eggs			1909–2010			Retail weight
		20	Fish and shellfish			1909–2010			Fish and shellfish (boneless weight)
		21	Coffee, tea, and cocoa			1910–2010			Coffee green bean equivalent + tea dry leaf equivalent + cocoa bean equivalent
		22	Legumes			1909–2010			Farm weight
PRODUCTION (P)		23	Hay Alfalfa	Acres harvested (ac hvt./year)	County	1965–2012	United States Department of Agriculture (USDA)	http://quickstats.nass.usda.gov/	SLB = Counties addition (Alamosa + Conejos + Costilla + Rio Grande + Saguache)
		24	Potatoes	Acres planted (ac plt./year)		1965–2011	(National Agricultural Statistics Service, 2013)		Oats estimated values: 2009=(2008 + 2007)/2; 2010=(2008 + 2009)/2
		25	Barley All			1956–2012			
		26	Wheat Spring excluding durum			1929–2012			
		27	Oats						

Ultimately, Fisher information sustainability index was calculated independently for each component of the SLB (e.g., demographic) (Table 2) using a window of two years (containing 8 time steps) and moving the computation one time step (i.e., 0.25 year) at a time. The preprocessed variables (i.e., normalized and interpolated) as well as the I result per component were filtered using a moving average of one year in order to smooth out fluctuations and focus on trends in the index. This I calculation has a twofold purpose: first to assess the trends in past data for each component of the system and second to obtain the information content on these trends. This information content represented by the I result from the preprocessed data is the external input that will be used to help avoid over fitting the model and serves as a means of bounding the forecast.

2.3.2. Artificial neural net forecasting

After the system is fully characterized and I has been computed to capture the patterns in each component, the I results are then used as an external input into the neural network. Fig. 2b provides flowchart of the neural net forecasting process. It shows the how the variables are preprocessed for each component, as in the characterization of the system (Fig. 1a). Fig. 2b also specifies the neural net design (i.e., training, validation and test) and application to forecast a baseline scenario for the SLB. This baseline scenario output (2011–2025) uses the I average from 1969 to 2010 as an external input to abound the projection (Fig. 2b).

In addition to variable normalization and interpolation, each time series was divided into three independent groups for cross-validation (Fig. 2b). Accordingly, data points in each variable were randomly sampled to train (i.e., 115 data points), validate (i.e., 25 data points), and test (i.e., 25 data points) the neural net architecture. The variables were modeled in groups by component of the system (Table 2). This data division helps the iterative process by providing a good number of net inputs, where the smallest set is the demographic component with 230 data points. Further, in order to avoid the opposite effect of having a very large data set, the food consumption component was divided into two groups: The first group contains six variables (i.e., Food consumption part 1: Fruit & vegetables, dairy products, flour & cereal products, caloric sweeteners, potatoes, and red meat) and 5 variables comprise the second group (Food consumption part 2: Added fats & oils, poultry, eggs, fish & shellfish, coffee tea & cocoa, and legumes). Therefore, a maximum of 990 data points were used for the food consumption component part 1.

Each component of the system (Table 2) was forecasted using the same DARX architecture (Fig. 3) and different delays were selected to get the best performance (Table 3). The result with the minimum squared error is presented for each component of the SLB regional system in Figs. 4–9. These figures are divided into two parts: the top part (a) of each graph contains the historical data (1969–2010) and forecasted (2011–2025) time series and the bottom part (b) has the external input (I) used to constrain this prediction. Note that in part b of each figure, there are two reference dashed lines ($I \pm 2SD(I)$) that help to determine whether the I is stable according to the criteria in Equation (2) and to identify deviations from normal behavior (Equation (3)).

2.3.2.1. Dynamic autoregressive artificial neural net with external input (DARX) architecture. The Matlab® software version *R2013a* and Neural Network Toolbox *8.0.1* were used to design, train, validate and test the architecture of the DARX used in this study (Fig. 3). This DARX architecture has one external input ($x(t)$), tapped delay lines (time window for calculation d) where the inputs ($y(t)$ and $x(t)$) are compared against a target (or desired output) to train (find by iteration weights (w) and biases (b)), validate and test the neural net. For the system under study, there are k inputs or variables ($y(t)$) describing each component of the San Luis Basin, Colorado region (e.g., $k=3$ for the environmental component, Table 2), 10 hidden layers of neurons with a positive linear transformation function, and k output results transformed linearly as well. For each component of the system (Table 2), DARX was tuned by changing the delay (d) from 4 to 12 years in order to achieve the best performance (minimum squared error), as it is shown in Table 3. The algorithm developed for this study is in Appendix B. There is no standardized rule to select the number of input variables or hidden layers for designing a neural net (Khashei and Bijari, 2012). As such, we employed the software default value of 10 for number of hidden layers and used this set-up to forecast each system component separately. The algorithm chosen for numerical optimization is the Levenberg–Marquardt (LM) method for training this artificial neural net (Ranganathan, 2004).

3. Results

3.1. Assessment of historical trends (1969–2010)

Fig. 4 shows the trajectory of the demographic characteristics of the system under study (Fig. 4a) and the I for this component (Fig. 4b). When looking at the plot of the time series data from 1969 to 2010, note that while both the personal income and population were essentially on an increasing trend, the I remained within a stable regime ($I \pm 2SD(I)$), except for a drop in I that occurred between 1994 and 1999. This minimum appears to correspond with a change in the trajectory of the demographic system described by population and personal income in the San Luis Basin. After that time, population seemed to taper and stabilize. Although, the maximum quinquennial growth rate for the SLB region was a 93% increase in Hinsdale's population (1971–1976), 1994 to 1999 was a common period of great growth for all counties, i.e., Hinsdale 31%, Mineral 24%, Saguache 16%, Costilla 10%, Rio Grande 9%, Conejos 7%, and Alamosa 6% (Figure D1). In contrast, the period 2004–2009 had a decreasing or steady trend in all counties, i.e., Mineral −14%, Costilla −5%, Saguache −4%, Rio Grande −4%, Conejos −3%, Hinsdale 3%, and Alamosa 0% (Figure D1). In the same fashion, personal income presented the lowest growth rate in the last years of study (2005–2010), i.e., Costilla 12%, Mineral 14%, Saguache 14%, Rio Grande 15%, Conejos 15%, and Alamosa 22%, and Hinsdale 27% (Figure D2).

Energy consumption in the system under study is mainly characterized by the use of fossil fuels (Fig. 5) and reflected a drastic change around 1980 to1982 (I minimum in Fig. 5b). This change in I during the beginning of the 80's echoed the period where there was the deepest decrease in the consumption of petroleum and natural gas, and a rapid increase in coal consumption (Fig. 5a). Similarly, the

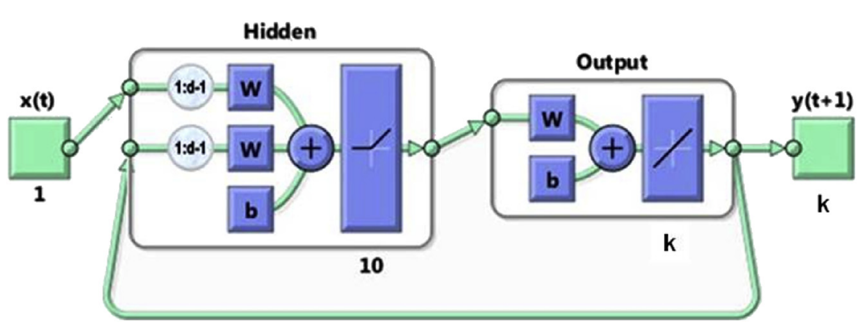

Fig. 3. DARX architecture. External input x(t): I calculated from y(t) (1970–2010) and I mean used to bound forecasted data (2011–2025), Input y(t): time series describing subcomponents of the system under study, hidden layers = 10 neurons (constant), delay: d changing to achieve best performance, training function updates weight and bias values according to Levenberg–Marquardt optimization rule and stop validation at a minimum mean squared error.

Table 3
Criteria selection for DARX. The best performance determines the number of years (delay) used to forecast subcomponents of the system, i.e., 7 years for demographic variables.

Delay (years)	Performance (minimum mean squared error)					
	Demographic	Environmental	Energy consumption	Food consumption		Production
				Part 1	Part 2	
	Variables 1-2	Variables 3-5	Variables 6-10	Variables 11-16	Variables 17-22	Variables 23-27
4	5.16E-02	3.70E-03	1.15E-02	3.30E-01	1.96E-02	8.84E+01
5	3.52E-06	7.08E-04	2.19E+04	4.80E-02	4.02E-04	1.60E+01
6	3.52E-06	6.14E-06	1.70E+01	9.33E-05	1.62E-04	8.44E-04
7	4.52E-07	8.08E+01	8.34E-06	2.17E-02	3.50E-03	3.73E-04
8	1.80E-03	8.72E-07	1.45E-04	1.75E-06	3.92E+02	3.90E-03
9	7.92E-07	5.44E-05	1.38E-04	5.95E-06	1.19E-04	2.37E-05
10	1.64E-05	6.55E-05	7.68E-06	1.02E-06	1.59E-05	3.39E-05
11	7.50E-07	3.69E-06	1.12E-05	3.94E-06	2.98E+20	4.01E-04
12	2.82E-06	1.45E-06	1.02E-05	9.76E-07	2.19E-06	9.54E-05

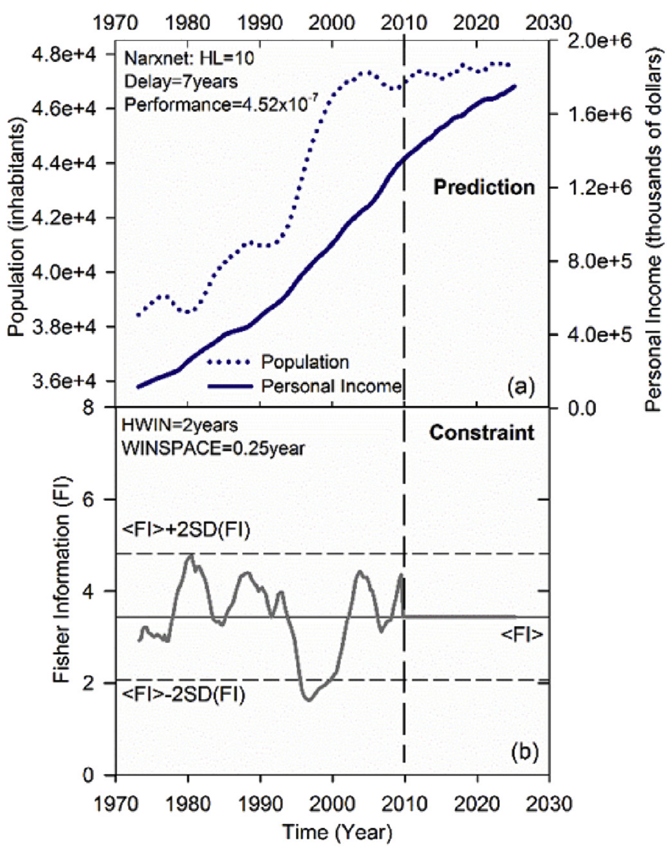

Fig. 4. Demographic.

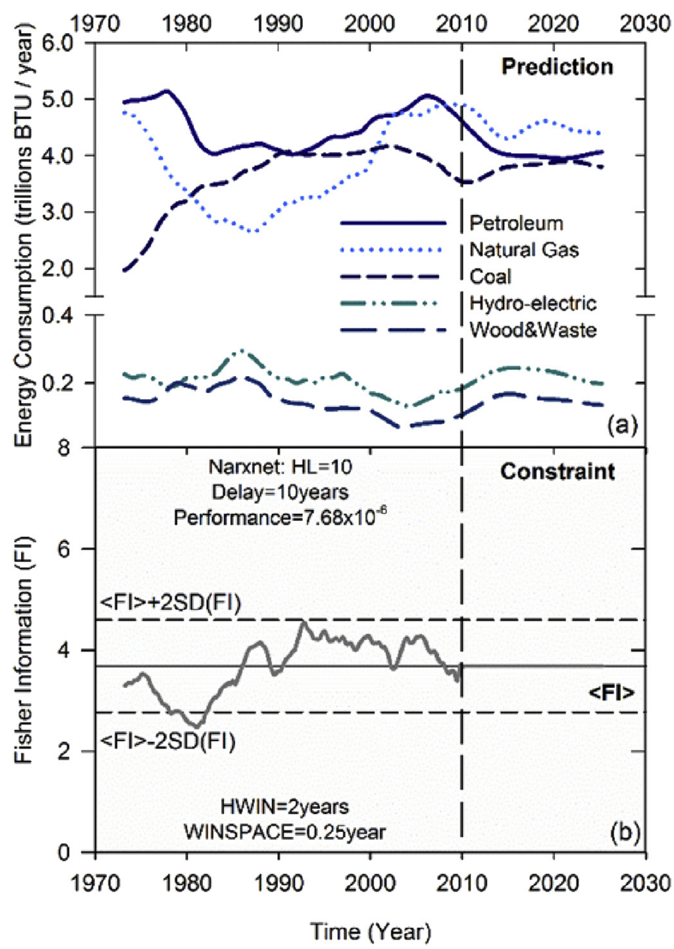

Fig. 5. Energy consumption.

environmental component of this system shows a *I* minimum between 1980 and 1982 (Fig. 6b) that coincides with relatively swift changes in both CO2 emissions (decrease) and precipitation (increase), which corresponds to the drop in *I* for energy consumption variables.

For the most widely available products by weight (>5 million lbs/year) such as, fruits and vegetables, dairy, flour, cereals, caloric sweeteners, potatoes and red meat, there was a rapid increase in their consumption around the years 1995 and 1997 (Fig. 7a). This change is represented by the *I* minima in Fig. 7b during that same period. While the food products that are consumed at a rate of less than 5 million lbs/year (e.g., eggs, fish, coffee, tea, cocoa and legumes) remained relatively steady during the period, there was a rapid increase in the added fats and oils, and poultry consumed which corresponded to a drop in *I* in 1999 (Fig. 8). Further, this shift in 1999 was preceded by the shift (1995–1997) in the greater consumed products (>5 million lbs/year) and that of the demographic component (1995) which reflected a similar dynamic trend.

Although there was quite a bit of change in the agricultural production variables (Fig. 9a), the *I* remained within two standard deviations of the mean, and it, therefore, remained stable (Fig. 9b). The time series trends and corresponding *I* provide a baseline scenario that is beneficial for planning purposes under conservative measures and was used to aid in developing and bounding the

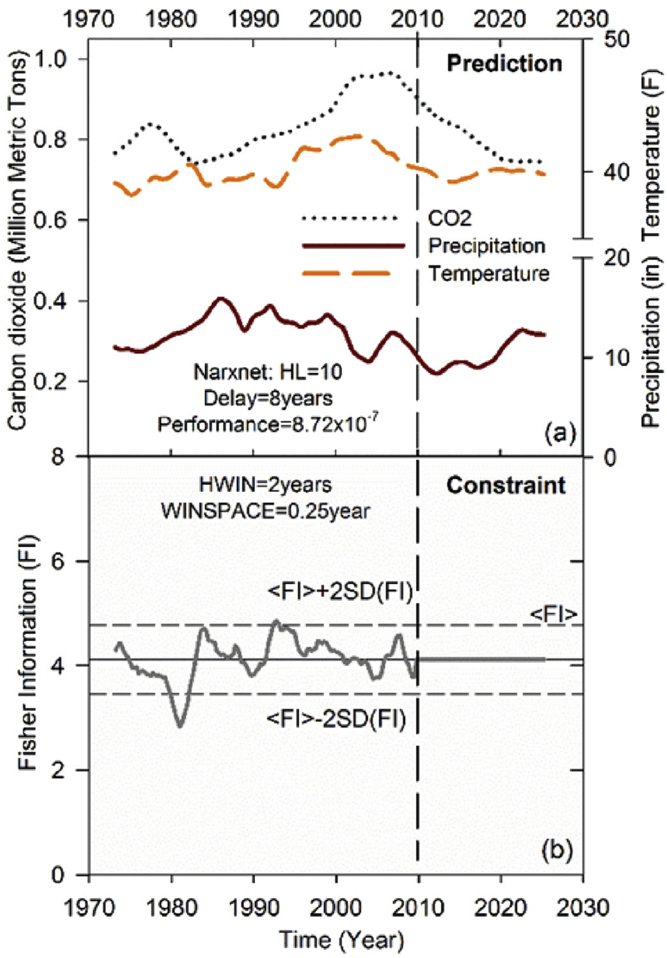

Fig. 6. Environmental.

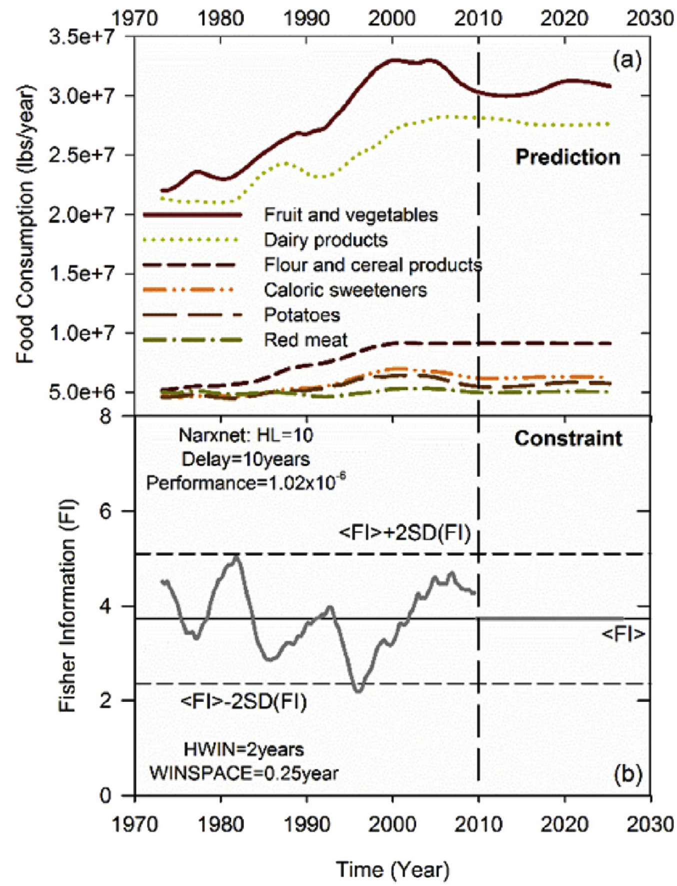

Fig. 7. Food consumption part 1.

projections from 2011 to 2025.

3.2. Evaluation of the projections (2011–2025)

Once, the neural network was trained, validated and tested, it was used to forecast the 27 time series from 2011 to 2025. As mentioned previously, these projections represent the baseline case such that the trends in the system variables are in line with historical patterns. The base case was simulated by using the mean *I* computed from each component of the system as an external constraint for DARX. Starting with the population data, note that there is a period of slow growth (1969–1990), a rapid transition to exponential growth (1995–1999), and a stabilization period (2000–2010). After which the prediction (2011–2025) maintains stability without drastic shifts. The forecast results in a modest 1% increase in the SLB population from 2011 to 2025 (Fig. 4a). Attempting to fit this variable to a polynomial (quadratic) would have resulted in a 16% increase (Figure D3) and therefore considered as an extraordinary event rather than an expected trend appropriate for a baseline scenario.

While the energy consumption in this area is not expected to suffer drastic changes, it is anticipated that natural gas consumption will exceed that of petroleum and coal (Fig. 5a). It is projected that there will be no major changes in temperature or precipitation. However, projections indicate that the carbon dioxide emissions will decrease to 1983 levels in the near future (Fig. 6a); reflecting a state of the environment towards a more sustainable trend. Figs. 7 and 8 show that food consumption will reach a plateau which is consistent with the projected population trends in the SLB. Likewise, the agricultural production in the San Luis Basin Region is expected to remain very close to its current level (Fig. 9).

In order to test whether the projections were consistent with a baseline scenario, *I* was computed for the projections of all 27 variables. Fig. 10 indicates that while the system tended to increase its dynamic order, the *I* of SLB region remained relatively stable (I±2SD(I)) from 2011 to 2025.

4. Discussion

Attempting to forecast a scenario is not a straightforward task. There are many challenges which inhibit this effort to include variable selection, data availability and forecasting approach. First, the depiction of reality brings intrinsic limitations that cascade into all steps of the process including the selection of variables describing a system, data availability and uncertainty, model development, simulation, calibration, validation, testing and performance (Bennett et al., 2013; Kelly et al., 2013; Laniak et al., 2013). In order to overcome disadvantages related to the data, we ensured that the best available information was used and the most representative scale possible (county was desired). When county level data was unavailable, the next larger system was selected to represent the SLB. Although this methodology is a data-driven statistical method, which was able to reproduce known values optimally (DARX was cross validated with the best performance), measurement uncertainty was unknown and therefore not used to define a lower and upper bounds for the base line scenario.

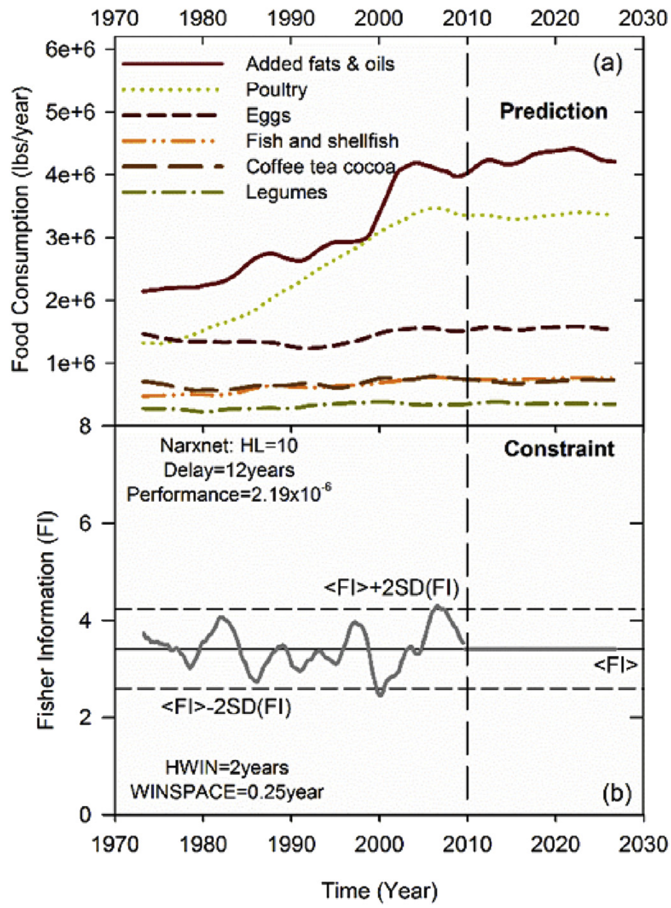

Fig. 8. Food consumption part 2.

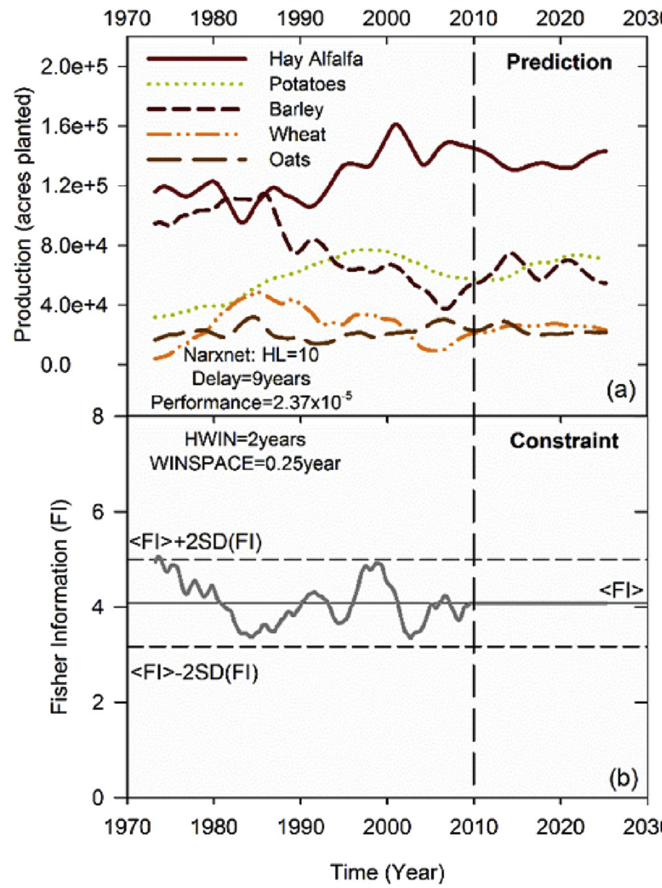

Fig. 9. Production.

However, assessment of the projections seems to indicate that the trends were maintained as expected, i.e., without drastic changes in the trajectory of the system. The methodology presented here demonstrates a data-driven tool to forecast variables describing a multi-dimensional system under an external constraint to ensure the maintenance of trends in the system variables.

A major strength and novelty of the forecasting methodology presented is the application of a sustainability index with a twofold purpose, system characterization and extrapolation constraint. First, this index was employed to describe historical trends on the demographic, environmental, consumption and production components of the SLB agricultural system. This characterization was used to ensure that the forecast did not have more information than the historical data used to develop the prediction. Second, when using a constraint (I average computed using historical data from 1969 to 2010) as the external input of a dynamic autoregressive neural network, the prediction maintained the time series trends and produce a business as usual scenario for the San Luis Basin, Colorado regional system. This baseline scenario can be used for regional planning and decision making regarding main sectors of the economy such as production and consumption of food and energy. Further, this method can also be applied to different systems characterized by similar variables or time series describing other social, economic, and environmental aspects. Future research about this forecasting approach could be focused on adjusting the external constraint to generate different scenarios that include uncertainty of the variables in order to define lower and upper bounds for the projections.

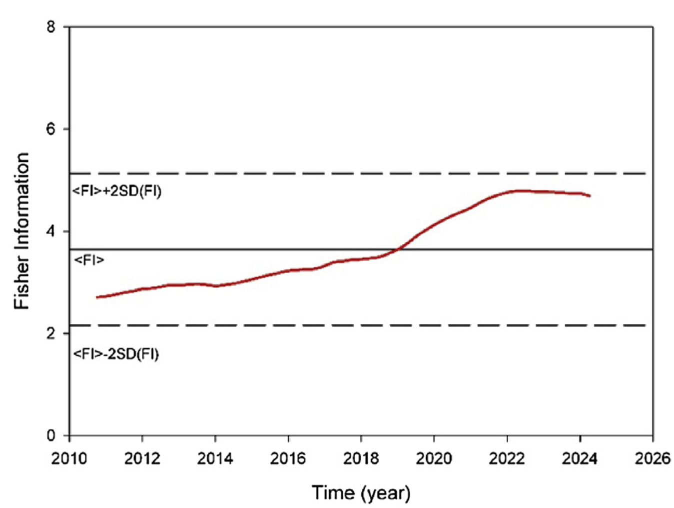

Fig. 10. Fisher Information for 27 forecasted variables (2011–2025).

4.1. The San Luis Basin region

The San Luis Basin region is a strategic partner of the State of Colorado and is prominent for producing potatoes and barley used to brew Coors and Budweiser beer, as well as, high-quality dairy hay (Colorado Office of Economic Development and International Trade (2014)). Hence, it can be influenced directly by Colorado's economic development, as well as, the economic cycles of the US.

Accordingly, national and state level trends (e.g., growth between the late-1960s and early-1980s, the economic recession and slowdown of the 1980s, the rebound in the 1990s, and stabilization since the beginning of the century) have impacted SLB region.

One example of these trends for the SLB is the average price per acre of land and buildings, which had the maximum growth between 1978 and 1982 (23% annual increase) that dropped to a −3% annual decrease from 1982 to 1987 remaining stable until 1992, and picking up in the 1990s with a 9% annual increase between 1992 and 2002, and a 4% annual increase from 2002 to 2007 (Figure D4).

Figs. 5–9 show the early-1980s recession for the SLB and two periods of growth, 1969–1979 and 1995–1999. This echoes Colorado's trends before and after the end of the oil and gas boom, the real state bust, and critical job losses in 1982 (Kendall, 2002). Another example of these dynamics during the 1980s recession can be seen in Fig. 5, where fossil fuels consumption in the SLB was the lowest; therefore, the levels of carbon dioxide had the slowest increase (Fig. 6). In contrast, Colorado's 1990s economic rebound (Lyons and Johnson, 2006) corresponds to the highest population growth rate for the SLB, which was 14% from 1990 to 1999, compared to only 6% increase for the previous decade, 2% in the 1970s, and just 1% from 2000 to 2009 (Fig. 4). The 1990s economic growth also reported higher consumption of food (Figs. 7 and 8), as well as, a moderate increase in agricultural production (Fig. 9) that by 2000 slowed and stabilized in the last 10 year under study.

For the period 2000–2010, farm employment decrease 9% and non-farm employment grew 11% indicating a swift shift from agricultural occupation to services such as, health care and social assistance, accommodation and food services, construction, professional, scientific, and technical services, arts, entertainment, and recreation (U.S. Bureau of Economic Analysis, 2012). The projected time series maintains this trend may be useful for decisions related to services in such areas as the Great Sand Dunes National Park and Preserve, the Rio Grande and San Juan National Forests, and at three National Wildlife Refuges, Alamosa, Monte Vista, and Sangre de Cristo, Rio Grande and the Conejos rivers for fishing, the Wolf Creek Ski Area, and Penitente Canyon climbing slopes (San Luis Valley Development Resources Group 2013).

5. Conclusion

An important need in sustainable environmental management is the ability to gage possible changes over time and to develop credible future scenarios on which decisions can be based. Scenarios are an important tool to get an insight into different possibilities towards a sustainable future (Boyko et al., 2012). Science-based metrics and indicators provide a means of tracking conditions. However, available data is often years out of date, and it, therefore, does not necessarily reflect the current status of the system (Heberling and Hopton, 2014). This makes real-time decision making and future planning difficult. While past performance does not ensure a known future, there is a wealth of knowledge on system behavior in historical data. Thus, past trends may prove beneficial in order to forecast baseline conditions based on system dynamics.

Integrated environmental models are used extensively across different fields in order to quantify and represent system behavior (Laniak et al., 2013). Such knowledge of a system can be employed to extrapolate characteristics with some confidence (Bennett et al., 2013). There are traditional methods used to model particular aspects of natural and human systems, such as population dynamics (Murray, 2002) and econometrics (Maddala, 1992) to name but a few. However, previous information about the variable relationships (e. g., dependency) and high quality and extensive time series are required for the application of these models to estimate future values. Thus, a more intuitive method is preferable when dealing with real data (Kelly et al., 2013). The artificial neural network approach is a data-driven mathematical model that identifies the underlying factors, which might influence the variables being forecasted (Maier et al., 2010). They have successfully been used to develop scenarios from projected time series (Behrang et al., 2011). In spite of their popularity and adaptability, neural nets face challenges in designing the network topology and over fitting of the model, which requires huge amounts of data for an accurate prediction (Haimi et al., 2013).

This research article presents a novel methodology that employs the computational power of an artificial neural network and an information conservation principle from Information Theory. The advantages of the method illustrated in the present work are statistical robustness, which avoids under and over fitting of the model, input flexibility for time series selection, and the possibility of projecting different scenarios. By applying this approach to the SLB, Colorado, we were able to demonstrate that the method provides a means of forecasting time series variables to develop a baseline scenario using I as a sustainability constraint to preserve trends in the system trajectory. Such an approach is useful in sustainable environmental management and helps to facilitate sustainable decisions about consumption and production in complex human systems.

Disclaimer

The views expressed herein are strictly the opinions of the authors and in no manner represent or reflect current or planned policy by the federal agencies. Mention of trade names or commercial products does not constitute endorsement or recommendation for use. The information and data presented in this product were obtained from sources that are believed to be reliable. However, in many cases the quality of the information or data was not documented by those sources; therefore, no claim is made regarding their quality.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.envsoft.2016.03.002.

References

Antanasijević, D., Pocajt, V., Popović, I., Redžić, N., Ristić, M., 2013. The forecasting of municipal waste generation using artificial neural networks and sustainability indicators. Sustain Sci. 8 (1), 37–46.

Beale, M.H., Hagan, M.T., Demuth, H.B., 2013. Multilayer Networks and Backpropagation Training. Collect and Prepare the Data, Neural Networks Toolbox User's Guide. The MathWorks, Inc., Natick, MA, p. 422.

Behrang, M.A., Assareh, E., Assari, M.R., Ghanbarzadeha, A., 2011. Using bees algorithm and artificial neural network to forecast world carbon dioxide emission. Energ Sources 33 (19), 1747–1759.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40 (0), 1–20.

Bond, A., Morrison-Saunders, A., 2010. Re-evaluating sustainability assessment: aligning the vision and the practice. Environ. Impact Asses 31 (1), 1–7.

Boyko, C.T., Gaterell, M.R., Barber, A.R.G., Brown, J., Bryson, J.R., Butler, D., Caputo, S., Caserio, S., Coles, R., Cooper, R., Davies, G., Farmani, R., Hale, J., Hales, A.C., Hewitt, C.N., Hunt, D.V.L., Jankovic, L., Jefferson, I., Leach, J.M., Lombardi, D.R., McacKenzie, A.R., Memon, F.A., Pugh, A.M., Sadler, J.P., Weingaertner, C., Whyatt, J.D., Rogers, C.D.F., 2012. Benchmarking sustainability in cities: the role of indicators and future scenarios. Glob. Environ. Chang. 22 (1), 245–254.

Bureau of Land Management-Colorado/San Luis Valley Field Office, 2014. San Luis Valley Detailed Land Status Map.

Campbell, D.E., Garmestani, A.S., 2012. An energy systems view of sustainability: emergy evaluation of the San Luis Basin, Colorado. J. Environ. Manag. 95 (1),

72–97.
Colorado Office of Economic Development & International Trade, 2014. Colorado blueprint. A Bottom-up Approach to Economic Development, p. 3. Denver, CO.
Darbellay, G.A., Slama, M., 2000. Forecasting the short-term demand for electricity: do neural networks stand a better chance. Int. J. Forecast. 16, 71–83.
Eason, T.N., Cabezas, H., 2012. Evaluating the sustainability of a regional system using fisher information in the San Luis Basin, Colorado. J. Environ. Manag. 94 (1), 41–49.
Economic Research Service, 2013. Food Availability per year.
Fath, B.D., Cabezas, H., Pawlowski, C.W., 2003. Regime changes in ecological systems: an information theory approach. J. Theor. Biol. 222 (4), 517–530.
Fisher, R.A., 1922. On the mathematical foundations of theoretical statistics. Philosophical Trans. R. Soc. Lond. 222, 309–368.
Frieden, B.R., 2004. Science from Fisher Information. A Unification. Cambridge University Press, New York, NY, USA.
Gonzalez-Mejia, A.M., Eason, T.N., Cabezas, H., Suidan, M.T., 2012. Assessing sustainability in real urban systems: the greater Cincinnati metropolitan area in Ohio, Kentucky, and Indiana. Environ. Sci. Technol. 46 (17), 9620–9629.
Gonzalez Mejia, A.M., Eason, T.N., Cabezas, H., Suidan, M., 2014. Social and economic sustainability of urban systems: comparative analysis of metropolitan statistical areas in Ohio, Usa. Sustain. Sci. 9 (2), 217–228.
Guresen, E., Kayakutlu, G., Daim, T.U., 2011. Using artificial neural network models in stock market index prediction. Expert Syst. Appl. 38 (8).
Haimi, H., Mulas, M., Corona, F., Vahalaa, R., 2013. Data-derived soft-sensors for biological wastewater treatment plants: an overview. Environ. Modell. Softw. 47, 88–107.
Heberling, M.T., Templeton, J.J., Wu, S., 2012. Green net regional product for the San Luis Basin, Colorado: an economic measure of regional sustainability. J. Environ. Manag. 111, 287–297.
Heberling, M.T., Hopton, M.E., 2014. Assessing sustainability when data availability limits real-time estimates: using near-time indicators to extend sustainability metrics. Clean. Techn Environ. Policy 16 (4), 739–748.
Hopton, M.E., Cabezas, H., Campbell, D., Eason, T., Garmestani, A.S., Heberling, M.T., Karunanithi, A.T., Templeton, J.J., White, D., Zanowick, M., 2010. Development of a multidisciplinary approach to assess regional sustainability. Int. J. Sust. Dev. World 17 (1), 48–56.
Hopton, M.E., White, D., 2012. A simplified ecological footprint at a regional scale. J. Environ. Manag. 111, 279–286.
Jain, A.K., Mao, J., Mohiuddin, K.M., 1996. Artificial neural networks: a tutorial. Computer 29 (3), 31–44.
Kankal, M., Akpınar, A., Kömürcü, M.İ., Özşahin, T.Ş., 2011. Modeling and forecasting of Turkey's energy consumption using socio-economic and demographic variables. Appl. Energy 88 (5), 1927–1939.
Karunanithi, A.T., Cabezas, H., Frieden, B.R., Pawlowski, C.W., 2008. Detection and assessment of ecosystem regime shifts from Fisher information. Ecol. Soc. 13 (1), 22.
Kates, R.W., 2011. What kind of a science is sustainability science. Proc. Natl. Acad. Sci. U. S. A. 108 (49), 19449–19450.
Kelly, R.A., Jakeman, A.J., Barreteau, O., Borsuk, M.E., ElSawah, S., Hamilton, S.H., Henriksen, H.J., Kuikka, S., Maier, H.R., Rizzoli, A.E., van Delden, H., Voinov, A.A., 2013. Selecting among five common modelling approaches for integrated environmental assessment and management. Environ. Modell. Softw. 47 (0), 159–181.
Kendall, W.D., 2002. A Brief Economic History of Colorado Prepared for the Demography Section. Colorado Department of Local Affairs by the Center for business and Economic Forecasting, Inc., Denver, CO, p. 27.
Khashei, M., Bijari, M., 2012. A new class of hybrid models for time series forecasting. Expert Syst. Appl. 39, 4344–4357.
Krogh, A., 2008. What are artificial neural networks. Nat. Biotechnol. 26 (2), 195–197.
Laniak, G.F., Olchin, G., Goodall, J., Voinov, A., Hill, M., Glynn, P., Whelan, G., Geller, G., Quinn, N., Blind, M., Peckham, S., Reaney, S., Gaber, N., Kennedy, R., Hughes, A., 2013. Integrated environmental modeling: a vision and roadmap for the future. Environ. Modell. Softw 39 (0), 3–23.
Lapin, L.L., 1975. Statistics : Meaning and Method Harcourt Brace Jovanovich. NY USA, New York.
Lyons, K., Johnson, N., 2006. Education and Investment, Not Tabor, Fueled Colorado's Economic Growth in 1990s. Center on Budget and policy Priorities, Washington, DC, p. 10.
Maddala, G.S., 1992. Introduction to Econometrics, third ed. John Wiley & Sons LTD.
Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. Environ. Modell. Softw. 25 (8), 891–909.
Mayer, A.L., 2008. Strengths and weaknesses of common sustainability indices for multidimensional systems. Environ. Int. 34 (2), 277–291.
Mayer, A.L., Pawlowski, C.W., Fath, B.D., Cabezas, H., 2007. Applications of Fisher information to the management of sustainable environmental systems. In: Frieden, B.R., Gatenby, R.A. (Eds.), Exploratory Data Analysis Using Fisher Information. Springer-Verlag London, pp. 217–244.
Murray, J.D., 2002. Models for Interacting Populations, Mathematical Biology I: an Introduction. Springer-Verlag, New York, pp. 79–118.
National Agricultural Statistics Service, 2013. Quick Stats. Agricultural District - Colorado (1969-2010).
National Climatic Data Center, 2013. U.S. Annual Climatological Summary. Climate Data Online: Search Tool by County (1969-2010).
Pawlowski, C.W., Fath, B.D., Mayer, A.L., Cabezas, H., 2005. Towards a sustainability index using information theory. Energy 30, 1221–1231.
Ranganathan, A., 2004. The Levenberg-marquardt Algorithm.
Tu, Y.-L., Chang, T.-J., Chen, C.-L., Chang, Y.-J., 2012. Estimation of monthly wind power outputs of WECS with limited record period using artificial neural networks. Energ Convers. Manage 59, 114–121.
U.S. Bureau of Economic Analysis, 2012. CA25N Total Full-time and Part-time Employment by NAICS Industry.
U.S. Bureau of Economic Analysis, 2013. Personal income, per capita personal income, and population (CA1–3) by County (1969-2010).
U.S. Census Bureau, 2013. State and County QuickFacts. Data Derived from Population Estimates, American Community Survey, Census of Population and Housing, State and County Housing Unit Estimates, County Business Patterns, Nonemployer Statistics, Economic Census, Survey of Business Owners, Building Permits.
U.S. Energy Information Administration, 2013a. Colorado Carbon Dioxide Emissions from Fossil Fuel Consumption (1980-2010).
U.S. Energy Information Administration, 2013b. Colorado State Profile and Energy Estimates (1969-2010).
U.S. Energy Information Administration, 2013c. Total Energy. Annual Energy Review. Environment. Table 11.1 Carbon Dioxide Emissions From Energy Consumption by Source, Selected Years, 1949-2011.
USEPA, 2010. In: Cabezas, H., Campbell, D., Eason, T., Garmenstani, A.S., Heberling, M.T., Hopton, M.E., Templeton, J.J., White, D., Zanowick, M., Sparks, R.T. (Eds.), San Luis Basin Sustainability Metrics Project: a Methodology for Evaluating Regional Sustainability. U.S. E.P.A., Cincinnati OH, p. 231.
Voyant, C., Muselli, M., Paoli, C., Nivet, M.-L., 2011. Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation. Energy 36 (1), 348–359.
Young II, W.A., Millie, D.F., Weckman, G.R., Anderson, J.S., Klarer, D.M., Fahnenstiel, G.L., 2011. Modelling net ecosystem metabolism with an artificial neural network and Bayesian belief network. Environ. Modell. Softw. 26 (10), 1199–1210.
Zaihong, Z., Zhiqiang, D., Rusch, K.A., 2012. Development of predictive models for determining enterococci levels at Gulf Coast beaches. Water Res. 46 (2), 465–474.