Environmental Modelling & Software xxx (2014) 1–12

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

# Analysing coastal ocean model outputs using competitive-learning pattern recognition techniques

Raymond N. Williams [a,b,*], Paulo A. de Souza Jr. [a], Emlyn M. Jones [c]

[a] CSIRO Computational Informatics, Hobart, Tasmania, Australia
[b] School of Engineering and ICT, University of Tasmania, Hobart, Tasmania, Australia
[c] CSIRO Marine and Atmospheric Research, Castray Esplanade, Hobart, Tasmania, Australia

ARTICLE INFO

*Article history:*
Received 2 December 2013
Received in revised form
28 February 2014
Accepted 5 March 2014
Available online xxx

*Keywords:*
Pattern recognition
Self-organizing maps
Ocean modelling
Data mining

ABSTRACT

To assist in interpreting the hydrodynamics of a complex coastal environment, a Self Organizing Map (SOM) has been constructed using output from a three-dimensional hydrodynamic model of the Huon-D'Entrecasteaux region in South-East Tasmania, over a one-year period. Interpretation of the SOM enabled nine characteristic or prototype states to be identified. As expected, the dominant forcing mechanisms were freshwater input via riverine discharge and input from oceanic waters. While these mechanisms are well understood, subtle features associated with the interaction of the two forcing mechanisms and the transitions between meta-stable states, were revealed by visualizing the SOM output. Further investigation was undertaken to determine how effective the SOM would be in identifying these prototype states given sensor data from a sensor network being designed for future deployment within the region. This research has demonstrated that SOM analysis can be a useful tool for identifying and interpreting patterns in large oceanographic datasets.



## 1. Introduction

As anthropogenic pressures on the coastal marine environment increase, managers and policy makers are formulating adaptive management strategies based upon quantitative scientific studies. Underpinning many of these studies is the integration of models and observations, using both qualitative (Wild-Allen et al., 2010) and quantitative (Ward et al., 2010) techniques. This has resulted in a phenomenal increase in the amount of observational data and model output that needs to be interpreted. Data assimilation techniques using Bayesian Inference (Parslow et al., 2013) and more approximate methods (Margvelashvili et al., 2013) allow for a robust statistical method to combine models and observational data. We are then confronted with the problem of analysing and interpreting the enormous datasets that are produced, in order to learn as much as possible from the data contained within them.

Traditional methods of interpreting model output through model-data assessments and comparisons are computationally demanding and difficult to visualise in 3D. Automated procedures utilising technological advances in machine learning and data-mining (Liu and Weisberg, 2005; Liu et al., 2009; Jin et al., 2010; Liu and Weisberg, 2011; Williams et al., 2012) are being adopted, yielding valuable information in identifying dynamics relevant to seasonal and climate time-scales.

As observing systems mature, and observational datasets are being received in near real-time from in situ sensor networks and satellite remote sensing, these datasets are being assimilated into coastal models. The observations and models then form the basis of coastal information systems that are being deployed in regional and national contexts in many coastal areas around the world. An example of one such system within Australia is the INFORMD system in South-East Tasmania (Margvelashvili et al., 2010), which is now being applied on a regional scale to the Great Barrier Reef Region, in a multi-organisation collaborative project entitled *eReefs* (www.barrierreef.org/OurProjects/eReefs.aspx). Similar systems are now operational in Europe (www.myocean.eu.org) and the US (www.ioos.noaa.gov). A common theme within these projects is the fact that the amount of data and model output being produced is increasing rapidly and better tools are required to interpret the model output and distil the information required to reveal the underlying dynamics of the systems of interest.

The overall goal of this study was to demonstrate the use of competitive-learning pattern recognition techniques to interpret

* Corresponding author. School of Engineering and ICT, University of Tasmania, Hobart, Tasmania, Australia.
*E-mail addresses:* Ray.Williams@csiro.au, R.Williams@utas.edu.au (R. N. Williams).

http://dx.doi.org/10.1016/j.envsoft.2014.03.001
1364-8152/

the dynamics in a coastal system where we presently have a good understanding of the different environmental conditions and their drivers.

In order to demonstrate this, we used an unsupervised competitive learning neural network algorithm called the Self-Organizing Map or SOM (Section 2.3) to identify prototype states exhibited by the Huon Estuary – D'Entrecasteaux Channel coastal region of South-East Tasmania over a one-year period from 1st March 2008 to 28th February 2009 (Section 2.3).

Once these prototype states had been identified, we trained a supervised competitive-learning neural network, the Learning Vector Quantization (LVQ) network (Section 2.4), to predict these prototype states from simulated real-time sensor data, in an effort to evaluate the effectiveness of different sensor network designs that may potentially be deployed across the Huon-D'Entrecasteaux coastal region (Fig. 1).

It is anticipated that strategic developments from this study will lead to the design and development of pattern recognition systems for use in larger more complex regions in which the underlying hydrodynamic factors involved are less well understood.

## 2. Methods

### 2.1. Computational modelling of ocean environments

While observations provided by a wide variety of ocean observing instruments can yield information about the present and past state of the system, they are generally unable to forecast future states. Therefore, computational models that are able to simulate alternative management scenarios and make predictions are a very important tool to develop effective adaptive management scenarios.

As part of a world-wide research effort to develop and refine ocean models, the Coastal Environmental Modelling team (www.emg.cmar.csiro.au) at CSIRO Marine and Atmospheric Research (CMAR), in Hobart, have developed the *Sparse Hydrodynamic Ocean Code (SHOC)* model. *SHOC* is a finite difference hydrodynamic model, based on an orthogonal curvilinear grid, capable of predicting three-dimensional distributions of temperature and salinity, as well as current velocity and direction, over scales ranging from estuaries to regional ocean domains, given appropriate inputs such as wind speed and direction, atmospheric pressure, surface heat fluxes and tides (Herzfeld et al., 2010).

The *SHOC* model has, over recent years, been applied to the south-east coastal region of Tasmania (Fig. 2). This coastal region includes two major estuaries, the Derwent estuary and the Huon estuary, linked by the D'Entrecasteaux Channel (Butler, 2006). It exhibits complex oceanographic behaviour because of the fact that two major ocean currents, the South Australia Current (SAC), which flows down the west coast of Tasmania, and the East Australia Current (EAC), which flows down its east coast, converge within the region. Also of interest is the contrast between the two major estuaries in the region; the Derwent estuary has suffered from significant urban pollution in the past, while the Huon estuary is in nearly pristine condition. In addition the D'Entrecasteaux Channel hosts a wide variety of recreational activities, including boating and fishing and also has a large and economically important aquaculture industry (Timms et al., 2010). Collectively, these very different coastal sub-regions provide an excellent test-bed for research into the sustainable management of coastal marine environments (Jones et al., 2012).

### 2.2. Pattern recognition in oceanography

Until recently, data collection within the oceanography field was difficult and the computational power available to oceanographers to analyse their data limited. This meant that oceanographic data analysis techniques were oriented toward deriving scarce information from small and homogeneous datasets using analysis techniques which were economical in their computational requirements. However the increasing volume and diversity of oceanographic data now becoming available threatens to overwhelm these techniques because they cannot easily discover new and unexpected patterns, trends and relationships hidden within large spatial datasets (Miller and Han, 2009).

Similar trends have been occurring in a number of scientific fields and, in response to this, new computational analysis techniques are being developed to extract features, trends and insights from the huge datasets now being produced by modern technologies, such as remote sensing satellites and ocean observatories. These techniques, some derived from traditional statistical methods and others from a branch of computer science called machine learning, automatically recognise new and unexpected patterns within data. These patterns can often be used subsequently by scientists to formulate or confirm hypotheses (Liu and Weisberg, 2011).

Statistical methods have been used extensively in the past to identify patterns in scientific data. They are mathematically rigorous and have been successfully applied to pattern recognition tasks in many scientific fields, including oceanography. However they generally assume that the data being analysed can be represented, at least approximately, using standard probability distributions. One popular example is time-series data analysis (Box et al., 2008).

The complexity and non-linearity of many of the processes underlying the environmental sciences mean that traditional statistical methods have limited application within these fields (Liu and Weisberg, 2005). Because of this, pattern recognition techniques employing methods derived from the field of machine learning are being evaluated by researchers to see what part they may play in environmental science research. They are non-linear methods and do not depend on the assumptions that traditional statistical techniques make about the datasets being analysed.

Within the machine learning field various pattern recognition techniques have been developed by computationally modelling some of the neurological processes known to underpin the pattern recognition capabilities of the human brain. Such models are called artificial neural networks because they mimic the behaviour of the networks of biological neurons, within our brain, which underlie our thinking processes (Dayhoff, 1990). Such techniques have an inherent ability to recognise non-linear relationships and are robust in the face of noisy and incomplete data, rendering them very suitable for use within the environmental sciences (Reusch et al., 2005; Haupt et al., 2009; Hsieh, 2009).

The most common type of neural network used within the environmental sciences to date has been a classification network called the back-propagation neural network (Wu et al., 2006), but over recent years a competitive-learning network called the *Self-Organizing Map (SOM)* has begun to generate significant interest (Cavaros, 2000; Tambouratzis and Tambouratzis, 2008; Shanmuganathan et al., 2006; Agarwal and Skupin, 2008; Kalteh et al., 2008; Morioka et al., 2010). SOMs have been applied successfully in coastal oceanography (Richardson et al., 2002, 2003; Demarcq et al., 2008; Jin et al., 2010; Iskandar, 2010) and have been used to analyse numerical ocean circulation model output (Iskandar et al., 2008; Liu et al., 2009). They have also been applied to a wide range of other environmental modelling applications in recent years, including waste water treatment plant operation (Dürrenmatt and Gujer, 2012; Liukkonen et al., 2013), reconstruction of past climatic conditions (Friedel, 2012), modelling crop evapotranspiration (Adeloye et al., 2012), groundwater transport (Friedel and Iwashita, 2013) and benzene pollution (Strebel et al., 2013), and constructing computational policy simulations for natural hazard mitigation (Samarasinghe and Strickert, 2013).

### 2.3. The Self-Organizing Map (SOM)

The SOM is an artificial neural network model (Kohonen, 1995) which can create a spatially organized representation of data closely resembling the ordered maps of cells observed within the cortex of the human brain, where cells near to each other respond to similar sensory inputs. The spatial proximity of these brain cells within the cortex reflect significant relationships between the input signals to which each responds. A SOM is a two-layer neural network (Fig. 3) with an input layer consisting of $n$ nodes (assuming that the network is intended to analyse $n$-dimensional datasets) and an output layer containing a specified number of nodes arranged (usually) in a two-dimensional grid structure. The number of nodes $M$ in the output layer is determined by the data analyst and depends on the level of detail required in the analysis. Each output node $i$ is connected to each input node via a weighted link and so associated with each output node is an $n$-dimensional weight vector $\mathbf{w}_i$ containing the weights of the links between this output node and all of the input nodes (Richardson et al., 2003). Adjacent nodes on the output grid are called neighbours.

Before the training process commences, the network weight vectors $\mathbf{w}_i$ are initialized, usually with random starting weights (*random initialisation*), although in

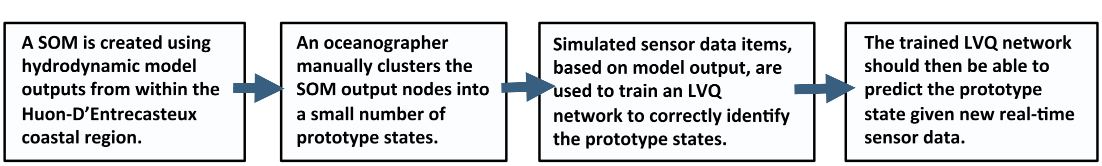


**Fig. 1.** Hybrid pattern recognition process that uses an unsupervised competitive-learning algorithm, the SOM (Section 2.3), to identify prototype hydrodynamic states and then trains an LVQ network (Section2.4), a supervised competitive-learning algorithm, to predict these prototype states from real-time sensor data.

Please cite this article in press as: Williams, R.N., et al., Analysing coastal ocean model outputs using competitive-learning pattern recognition techniques, Environmental Modelling & Software (2014), http://dx.doi.org/10.1016/j.envsoft.2014.03.001