Environmental Modelling & Software 93 (2017) 193–208

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

A hybrid empirical-Bayesian artificial neural network model of salinity in the San Francisco Bay-Delta estuary

John S. Rath [a], Paul H. Hutton [b], Limin Chen [a], Sujoy B. Roy [a, *]

[a] Tetra Tech, Lafayette, CA, United States
[b] Metropolitan Water District of Southern California, Sacramento, CA, United States

ARTICLE INFO

Article history:
Received 1 September 2016
Accepted 22 March 2017
Available online 30 March 2017

Keywords:
Salinity
Artificial neural network
Structural validation
Bayesian parameter estimation

ABSTRACT

This paper reports the refinement of a published empirical model of salinity in the San Francisco Bay-Delta estuary by integration with a Bayesian artificial neural network (ANN) model and incorporation of additional inputs. Performance goals established for the resulting hybrid model are based on the quality of fit to observed data (replicative and predictive validation) as well as sensitivity when compared with *a priori* knowledge of system behavior (structural validation). ANN model parameters were constrained to provide plausible sensitivity to coastal water level, a key input introduced in the hybrid formulation. In addition to representing observed data better than the underlying empirical model while meeting structural validation goals, the hybrid model allows for characterization of prediction uncertainty. This work demonstrates a real-world application of a general approach— integration of a pre-existing model with a Bayesian ANN constrained by knowledge of system behavior—that has broad application for environmental modeling.

1. Introduction

Artificial neural networks (ANNs) have seen growing application in the modeling of water resources and water quality processes over the past two decades. Reviews by Maier and Dandy (1996), Maier et al. (2010), and Wu et al. (2014) have documented more than 300 publications on the field, and readers are referred to these papers for a broad discussion of potential applications. ANN models are particularly attractive because they can be developed using standardized approaches even where mechanistic representations of a system are unavailable, prohibitively expensive computationally, or are too complex to formulate. As empirical formulations, ANN models are based on data that can be directly observed or simulated, as opposed to abstract processes that are embodied in mechanistic models. To develop such models, sufficient data must be available on the dependent variable of interest as well as independent variables known to affect the dependent variable, and appropriate software must be utilized to encapsulate these relationships through a data fitting (or calibration) process. This data fitting process is commonly referred to as "training", during which the adjustable parameters of an ANN model are estimated using error-minimization algorithms. Data are generally partitioned for training, with a fraction set aside for model validation. A successful model fit matches both the training and validation data. These steps are common to most modeling efforts, whether ANN-based or mechanistically-based, and have been termed "replicative" and "predictive" validity, respectively (Wu et al., 2014).

With the growing maturity of ANN applications in the literature, it has become clear that the "black-box" model relationship between inputs and outputs embodied in ANNs may not adequately represent the physical system being modeled (Jain et al., 2004; Kingston et al., 2005a). Thus, a trained and validated ANN model may fit the aggregate response to multiple inputs well, even though the sensitivity to a specific input is not physically meaningful, or in some cases, not physically plausible. The condition of representing inputs and outputs in a manner that is physically plausible, given an *a priori* understanding of a system, is termed "structural" validity (Wu et al., 2014). For a model to be successful under a wide range of future conditions, both predictive and structural validity are necessary prerequisites and have been the focus of a small subset of recent ANN applications (Jain et al., 2004; Kingston et al., 2005a; Jain and Kumar, 2009). Assessment of predictive validity is often the main criterion in the development of ANN models, as noted in a review of recent literature on water quality-related models by Wu

* Corresponding author.
E-mail address: sujoy.roy@tetratech.com (S.B. Roy).

http://dx.doi.org/10.1016/j.envsoft.2017.03.022
1364-8152/

et al. (2014). Of 99 ANN models evaluated by the authors, only 15 considered structural validity as a criterion.

The focus of this work is on the development of predictively and structurally valid ANN models for representing salinity behavior in the San Francisco Bay-Delta estuary, with performance better than existing models. The San Francisco Bay-Delta is the largest freshwater estuary on the Pacific Coast of the United States. The management (and therefore accurate prediction) of salinity in this region is of great consequence; it is tied to the health of a highly diverse ecosystem in the estuary and to the management of freshwater withdrawals that support the needs or more than 20 million people and 3 million acres of irrigated land in California, thus supporting one of the largest economies in the world (Luoma et al., 2015; Delta Plan, 2013). Salinity behavior in an environment such as the San Francisco Bay-Delta estuary is highly dynamic, with freshwater flow and tidal influences having time varying behavior on scales of hours to months (Schoellhamer, 2000; Monismith et al., 2002). A variety of empirical and mechanistic modeling tools have been developed to predict salinity behavior in the estuary for management decision making. Despite the availability of these tools, there remains an ongoing need for robust modeling approaches that can be applied with rapid turnaround times; the ANN modeling framework presented here is intended to meet this specific need. This work is motivated by the general utility of ANNs in modeling complex problems in the water resources domain, as well as the specific utility of ANN applications related to salinity intrusion in the San Francisco Bay-Delta estuary (Finch and Sandhu, 1995; Seneviratne et al., 2008; Rajkumar and Johnson, 2001; Chung and Seneviratne, 2009) and other estuaries (Maier and Dandy, 1996; Bowden et al., 2002, 2005; Huang and Foo, 2002; Conrads et al., 2006).

The model structure presented in this paper is a hybrid approach that integrates neural networks within an existing empirical modeling framework. Hybrid modeling approaches may achieve better results than using a single modeling approach (Maier et al., 2010) and can be used to provide some degree of structural validity in a modeling framework. Hybrid systems are an active area of research and Maier et al. (2010) see this as a maturation of the ANN methodology where their strengths are best used in conjunction with existing modeling approaches in the water resources domain. In this work, we utilized Bayesian inference—the application of Bayes' rule to obtain probabilistic estimates of parameters in statistical models—to calculate hybrid ANN model parameters; this approach allowed predictions to account for uncertainty in the model parameters and to improve structural validity by formally incorporating prior knowledge into the model through the use of prior distributions. This approach has been used with success elsewhere (Kingston et al., 2005b; Gelman et al., 2013; Humphrey et al., 2016). The consideration of structural validity in this work is of benefit by allowing for more robust model predictions; it also serves a broader purpose of addressing well-founded skepticism of ANN models based on their "black-box" character. This approach has the potential to improve overall model performance (e.g. quality of fits, incorporation of prediction uncertainty, and consideration of additional inputs in a structurally valid framework) compared to existing modeling frameworks for salinity in the San Francisco Bay-Delta estuary. Specific elements of this real-world application also hold potential in other environmental domains where ANN models are not targeted as a replacement for a pre-existing empirical or mechanistic model, but rather are used in tandem to enhance the performance of the resulting hybrid.

2. Background

Salinity management in estuaries is a subject of interest in basins where a significant amount of freshwater is diverted (Alber, 2002). Examples of similar well-studied systems occur globally and include the Murray-Darling River basin in Australia (Murray Darling Basin Ministerial Council, 1999), estuaries in the Southeastern United States along the Gulf of Mexico and Atlantic coast (Sklar and Browder, 1998; Reinert and Peterson, 2008), the Mekong River Delta in Vietnam (Dat et al., 2011), and the Guadalquivir River estuary in Spain (Fernández-Delgado et al., 2007).

Freshwater flow through the northern reach of the San Francisco Bay-Delta estuary, specifically the Suisun Bay and the western Delta of the Sacramento and San Joaquin Rivers (Fig. 1), is a function of seasonally and annually varying runoff as well as anthropogenic influences such as upstream reservoir releases and freshwater diversions. Over the past two decades, the balance between environmental and water export needs has been managed through the location of the low salinity zone, which is correlated with various aquatic species distributions (Jassby et al., 1995). Operationally, the low salinity zone is defined as the position of the 2 parts per thousand bottom salinity isohaline, termed X2. Under current regulations, it is interpolated as an equivalent surface salinity from fixed monitoring stations near the surface and reported as a distance from Golden Gate Bridge in kilometers (Fig. 1). Besides X2 position, which is largely driven by habitat considerations, salinity compliance is also maintained at several discrete locations to protect municipal and agricultural beneficial uses (CSWRCB, 2006). For example, agricultural salinity standards are maintained in the western Delta at Emmaton and Jersey Point (Fig. 1).

The basic conceptual model of freshwater-saltwater mixing in estuaries with seasonally varying flow patterns such as the San Francisco Bay-Delta estuary is as follows: freshwater flows repel salinity downstream (seaward) across a mixing zone (with longitudinal and vertical gradients) and saltwater intrudes upstream (landward) during periods of low freshwater flow. The extent of the salinity gradients varies with tides on hourly to daily time scales and varies with freshwater flows on daily to seasonal time scales. Salinity management in the estuary is primarily concerned with daily, fortnightly and seasonal variability; thus, salinity variation at sub-daily time scales is beyond the scope of this work.

Over the past two decades, a variety of modeling frameworks have been developed and applied to the prediction of salinity intrusion in the San Francisco Bay-Delta estuary. These frameworks range from simple empirical-statistical models to complex three-dimensional hydrodynamic models. Several simple empirical models are available to predict X2 position and/or salinity at fixed locations in the estuary as a function of flow on a daily time scale (Denton, 1993; Jassby et al., 1995; Monismith et al., 2002; MacWilliams et al., 2015; Hutton et al., 2015). The Delta Salinity Gradient (DSG) model proposed by Hutton et al. (2015) integrates the approaches used by Monismith et al. (2002) for X2 position prediction and Denton (1993) for fixed location salinity prediction as follows:

$$X2(t) = \Phi_1 * G(t)^{\Phi_2} \tag{1}$$

where G(t) is antecedent outflow and Φ_1 and Φ_2 are empirically determined constants. Antecedent outflow (Denton, 1993) is defined by the following routing function similar to one proposed by Harder (1977):