Journal of Hydrology 531 (2015) 187–197

Contents lists available at ScienceDirect

Journal of Hydrology

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

Data-driven models of groundwater salinization in coastal plains

G. Felisa [a], V. Ciriello [a], M. Antonellini [b], V. Di Federico [a], D.M. Tartakovsky [c,*]

[a] Department of Civil, Chemical, Environmental, and Materials Engineering, University of Bologna, Italy
[b] Department of Biological, Geological, and Environmental Sciences, University of Bologna, Italy
[c] Department of Mechanical and Aerospace Engineering, University of California, San Diego, USA

ARTICLE INFO

Article history:
Available online 5 August 2015

Keywords:
Statistical model
Aquifer management
Time series
Data analysis

SUMMARY

Salinization of shallow coastal aquifers is particularly critical for ecosystems and agricultural activities. Management of such aquifers is an open challenge, because predictive models, on which science-based decisions are to be made, often fail to capture the complexity of relevant natural and anthropogenic processes. Complicating matters further is the sparsity of hydrologic and geochemical data that are required to parameterize spatially distributed models of flow and transport. These limitations often undermine the veracity of modeling predictions and raise the question of their utility. As an alternative, we employ data-driven statistical approaches to investigate the underlying mechanisms of groundwater salinization in low coastal plains. A time-series analysis and auto-regressive moving average models allow us to establish dynamic relations between key hydrogeological variables of interest. The approach is applied to the data collected at the phreatic coastal aquifer of Ravenna, Italy. We show that, even in absence of long time series, this approach succeeds in capturing the behavior of this complex system, and provides the basis for making predictions and decisions.

1. Introduction

Reclaimed low coastal plains are among the most populated areas in the world. They are extensively exploited for a variety of purposes, including urban settlements, maritime and industrial infrastructure, agriculture and, not least, recreation and tourism (Small and Nicholls, 2003; McGranahan et al., 2007). Among diverse environments typical of coastal plains, active dune systems, paleo-dunes, and coastal forests are critical for providing protection from storm surge, beach erosion and sea spray, which can harm crops in coastal farmland or damage local structures and infrastructure (Gambolati et al., 2002; Feagin et al., 2005; Armaroli et al., 2012). Furthermore, these environments, together with coastal wetlands, are often the only residual natural areas in the coastal zone. As such, they are extremely valuable in terms of ecosystem services, tourism, and preservation of the historical heritage (Barbier et al., 2011).

Groundwater salinization in coastal areas is a pressing issue (Custodio, 2010), which is virtually certain to become worse in the future as a result of climate change and sea level rise (Oude Essink and Kooi, 2012; Mollema et al., 2012; Mollema et al., 2013). Aquifer salinization impacts coastal ecosystems, such as marshes, wetlands, and forests, since vegetation species richness and biodiversity are very sensitive to variations in water salinity (Antonellini and Mollema, 2010). It also affects soil quality and, therefore, crop productivity (Pitman and Läuchli, 2002).

Water resources management in low coastal plains is a challenging task, because the phreatic surface typically needs to be artificially controlled due to low topography and reclamation (Grootjans et al., 1998; Antonellini et al., 2008; De Louw et al., 2010; Oude Essink et al., 2010). Drainage is required to allow agriculture, prevent flooding of low-lands, and ensure dry ground conditions for coastal forests. At the same time, drainage contributes to water salinization by increasing saltwater intrusion (Antonellini and Mollema, 2010). Complex feedback mechanisms, which control ecosystems health in coastal areas, further complicate both management and modeling of low coastal planes (Giambastiani et al., 2014).

Characterization of groundwater salinization relies on the description of a series of driving processes such as aquifer recharge dynamics, sea water encroachment along rivers and channels, land reclamation drainage, water pumping from wells, upwelling of connate water from the bottom of the aquifer, evapotranspiration, and natural and anthropogenic land subsidence (Bear et al., 1999; Post et al., 2003; Antonellini et al., 2008). Analytical and numerical models (e.g., Bear et al., 1999; Cheng et al., 2001; Langevin et al., 2007) are usually employed to identify driving mechanisms and their contributions to the overall process of groundwater salinization. Such methods work well when either the significant factors

* Corresponding author.
E-mail address: dmt@ucsd.edu (D.M. Tartakovsky).

http://dx.doi.org/10.1016/j.jhydrol.2015.07.045
0022-1694/

controlling the process are few and constrained (Bear et al., 1999) or the amount of data available is extensive (e.g., Post et al., 2003; Oude Essink et al., 2010).

We present a case study in which these conditions are not satisfied and traditional modeling approaches fail to accurately describe the process of groundwater salinization (Antonellini et al., 2015). We consider the phreatic coastal aquifer of Ravenna (Italy), where multiple factors control the system behavior and the monitoring network is sparse. Rather than using under-parameterized flow and transport models, we apply statistical approaches to identify the underlying mechanisms of the salinization process and provide a basis for building predictive models. In particular, we employ time-series analyses (Hamilton, 1994, and references therein) to interpret the hydrological data collected in the study area. Auto-regressive moving average (ARMA) models (Box and Jenkins, 1976) are used to capture the stochastic behavior (temporal fluctuations) of the hydrological variables of interest and to investigate the relationships among them.

This article is arranged as follows. In Section 2, we provide a description of the study area. Section 3 presents the statistical methodology employed to model the nonlinear dynamics of a low coastal plane. Section 4 contains a discussion of the modeling results and their implications for the aquifer management. A brief summary of our study is presented in Section 5.

2. Study area

We focus on the low coastal plain of the Po river in the south of Ravenna (Italy), which is adjacent to the North Adriatic sea. The area stretches from Lido di Dante, in the north, to Lido di Classe, in the south, and extends from the shoreline westward for a few kilometers inland (Fig. 1). The area's geomorphology consists of a row of active dunes in the east, covered by halophilic bushy and grass vegetation, adjacent to multiple rows of paleo-dunes inland, covered by pine trees of the species *Pinus Pinaster* (Antonellini and Mollema, 2010). The paleo-dunes, on which the coastal pine forest grows, is drained by two ditches parallel to the coast. Farmland extends to the east, past the last row of paleo-dunes.

Alongshore sediment dispersal and land reclamation in the last 150–200 years has led to coastline progradation of up to 5 km. This induced switching from a brackish lagoonal setting to a continental one (Ciabatti, 1968; Veggiani, 1974; Stefani and Vincenzi, 2005). In recent years, beach erosion has reversed this trend as a result of a decreased sediment supply by rivers and land subsidence (Bondesan et al., 1995). Natural land subsidence, due to differential compaction of Pliocene and Pleistocene sediments, is about 2–3 mm/year. The current average anthropogenic subsidence rate, due to water and gas withdrawal, is on the order of 3 mm/year, with peaks in the study area of up to 15 mm/year. During the second half of the past century, these rates have been higher, reaching the maximum of abound 110 mm/year (Baú et al., 2000; Teatini et al., 2006).

The low topography, which reaches in some places 1 m below sea level, requires a land reclamation drainage system. A dense network of channels organized in mechanical drainage basins is deployed to lower the water table. The reclaimed water, in each drainage basin, is routed to a cluster of pumping machines and then uplifted into a channel draining into the sea (Stefani and Vincenzi, 2005). Coastline progradation and land reclamation have led to freshening of the groundwater in the upper permeable sedimentary sequence below the study area during the past 150–200 years; nevertheless, this trend has been reversed recently in some parts of the aquifer (Antonellini et al., 2015).

Shallow sediment deposition (less than 150 m from the surface) in the coastal area of Ravenna has been controlled by two transgression–regression cycles (Amorosi et al., 2004) that formed a sequence of sand and silty-clay bodies. The upper 20–30 m form

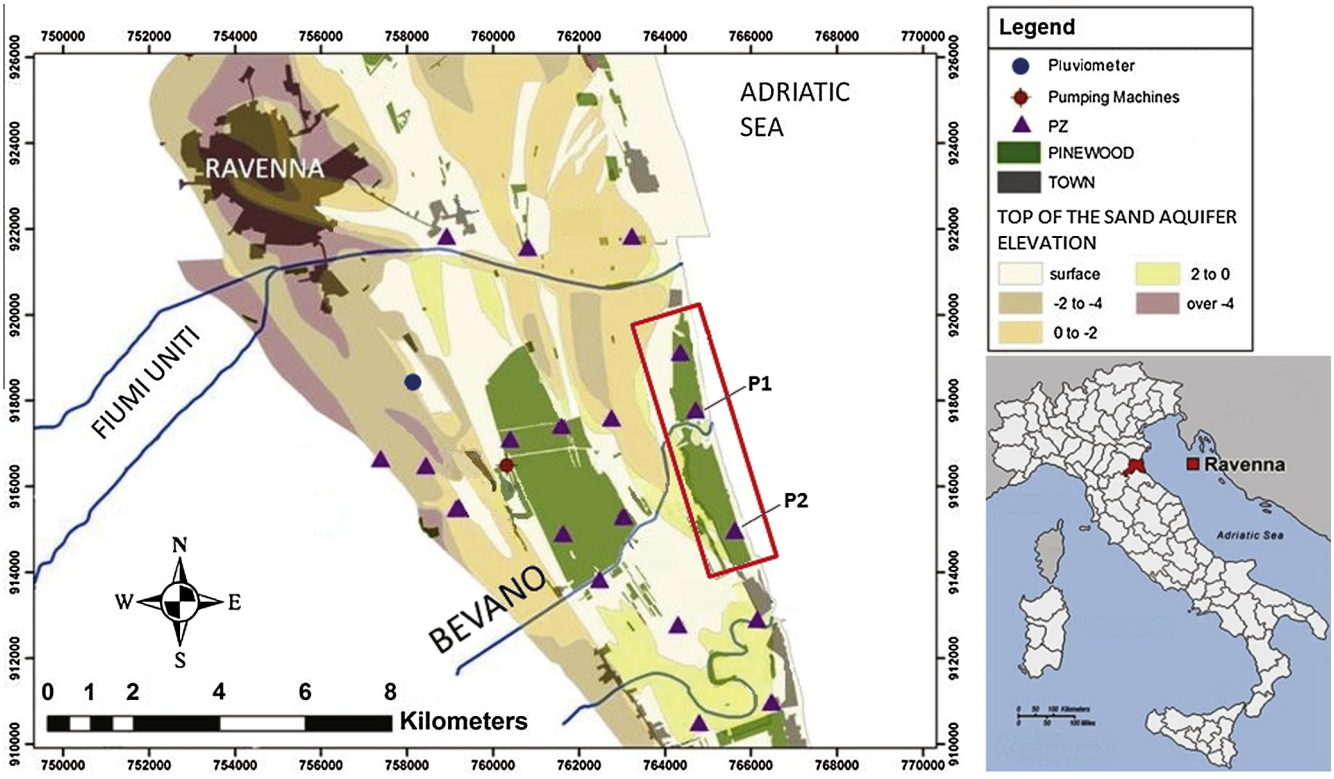

Fig. 1. Coastal region of Ravenna (Italy), with the red box indicating a study area comprising a coastal pine forest. Also shown are the piezometric and pluviometric monitoring networks and the pumping station considered in the analysis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

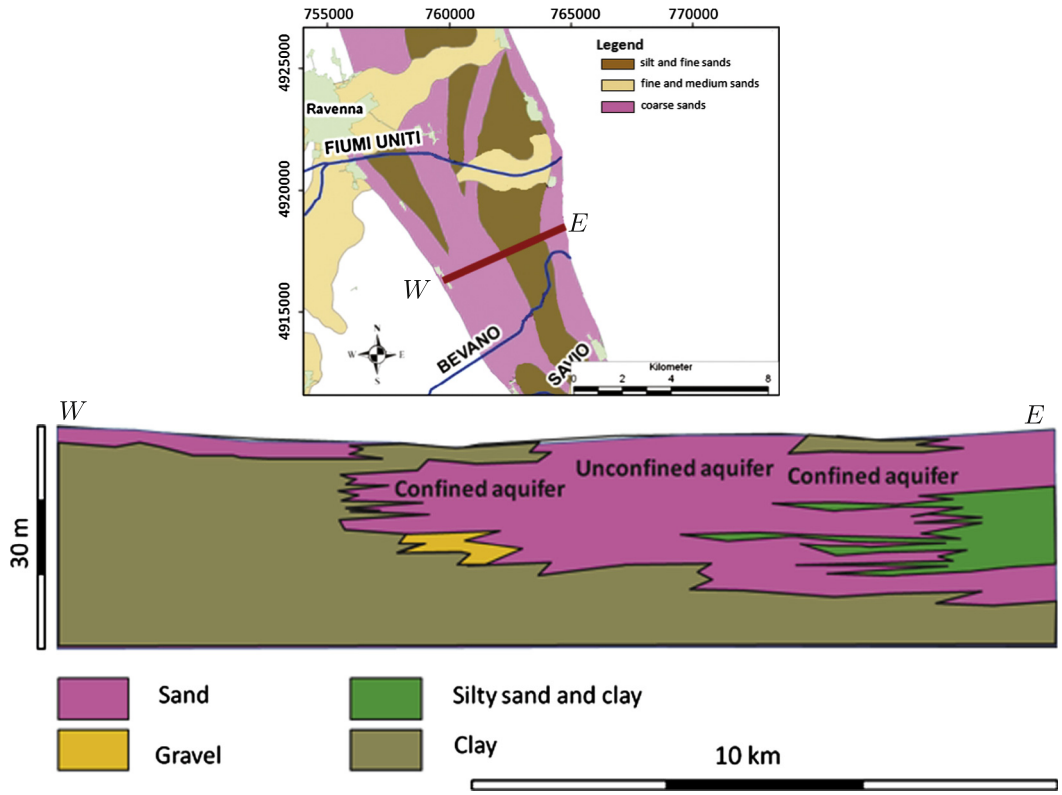

Fig. 2. Geologic map of the area south-east of Ravenna (modified from Regione Emilia-Romagna, http://ambiente.regione.emilia-romagna.it/geologia/cartografia/webgis-banchedati/webgis). The NE-SW-oriented red line represents the trace of the geologic cross-section.

the shallow coastal aquifer. They were deposited during the last sea level high stand (Marchesini et al., 2000). Most of the shallow aquifer is made up by beach bar sands. In the northern part of the study area, north of the Bevano River, some silty-clay deposits belonging to a distal delta sedimentary environment are found at an average depth of 10 m and have a thickness up to 15 m (Fig. 2). These fine-grained deposits disappear in the southward direction. It is important to note that the sands of the aquifer are often covered, at the surface, by thin layers of continental clay up to 1.5 m in thickness. They confine the surface aquifer along belts parallel to the coast (Amorosi et al., 2005).

The shallow coastal aquifer contains scattered freshwater lenses floating on top of brackish and saline water. The aquifer is recharged only via rainfall infiltration and excess irrigation water (Antonellini et al., 2008; Mollema et al., 2013). The origin of groundwater salinization in this area has been studied by Ulazzi et al. (2008), who observed conflicting relationships among the elevation of the water table, the health of the coastal pine forest, and the groundwater salt load. Marconi et al. (2011) and Mollema et al. (2013), by studying the hydro-geochemical characteristics of ground- and surface-water, established that a salinization trend in the aquifer is ongoing and controlled by seasonal variations in recharge. Groundwater salinization is mainly caused by two processes: (i) saltwater intrusion, because of the strong hydraulic gradients landwards and (ii) upwelling of Holocene brackish and salty water from the bottom of the aquifer, where the water table is below the sea level (Mollema et al., 2013). Factors, such as land subsidence, land use, drainage and seawater encroachment along rivers and channels, enhance these two processes (Giambastiani et al., 2007; Antonellini et al., 2008; Mollema et al., 2012; Mollema et al., 2013).

Given the hydrologic system's complexity and the data paucity, both construction of a reliable physics-based model and its parameterization are problematic. Instead, we employ statistical models that rely on the data collected at a series of piezometers, which have been installed in the last years to monitor the dynamics of the phreatic surface and the salinity of the shallow aquifer. These models are used to analyze the coastal plane area occupied by pine forest (Fig. 1).

3. Methodology

A hydrological time series may be seen as a single realization of a stochastic process. It can be modeled as the sum of (i) some deterministic components, which generate a systematic pattern, and (ii) a stochastic component (random noise), which represent the stationary counterpart of the hydrological process. The stationary part of a time series incapsulates the random nature of the selected process and is of fundamental importance for quantification of predictive uncertainty (Hamilton, 1994).

Auto-regressive integrated moving average (ARIMA) models remove the non-stationary component of a (hydrologic) time series by differentiating the original data (Box and Jenkins, 1976; Shahin et al., 1993; Shumway and Stoffe, 2005). If a time series is stationary, an auto-regressive moving average (ARMA) representation may be directly applied. These models assume that a phenomenon under consideration can be treated as a linear stochastic process. They have been employed to analyze hydrological time series, especially at the monthly scale (Wang et al., 2014, and references therein).

In the following, the main concepts related to these models are described and extended to the case of auto-regressive distributed lag (ARDL) models. The latter allow descriptions of processes, which involve multiple independent variables (time series) and are used to capture the influence of external stresses on a system's behavior e.g., (Almon, 1965; Davidson et al., 1978; Shin and Pesaran, 1999; De Boef and Keele, 2008; Beck and Katz, 2011).

3.1. ARMA models

Let us consider a stationary zero-mean time series consisting of n measurements y_t of a process $y(t)$, i.e., $\{y_t\} = (y_1, \ldots, y_n)$. An ARMA (α, β) representation of the time series $\{y_t\}$ is (Box and Jenkins, 1976)

$$y_t = \sum_{i=1}^{\alpha} \phi_i y_{t-i} + \sum_{j=1}^{\beta} \theta_j \xi_{t-j} + \xi_t, \tag{1}$$

where α and β represent the orders of the auto-regressive (AR) and moving average (MA) parts, respectively; ϕ_i and θ_j are model coefficients; and ξ_t denotes a white noise process. Eq. (1) is rewritten in terms of the lag operator $L^n y_t \equiv y_{t-n}$ as

$$\{\phi \cdot L\} y_t = \{\theta \cdot L\} \xi_t, \tag{2}$$

where $\{\phi \cdot L\} \equiv 1 - \phi_1 L - \ldots - \phi_\alpha L^\alpha$ and $\{\theta \cdot L\} \equiv 1 + \theta_1 L + \ldots + \theta_\beta L^\beta$ are the AR and MA polynomials, respectively.

If $\{y_t\}$ is a non-stationary time series, a way to obtain a stationary time series $\{\hat{y}_t\}$ is to differentiate $\{y_t\}$, i.e., (Box and Jenkins, 1976)

$$\hat{y}_t = (1 - L)^d y_t, \tag{3}$$

where d is the number of differences required to render the series $\{y_t\}$ stationary. For example, $d = 1$ yields $\hat{y}_t = (1 - L)^1 y_t = y_t - L y_t = y_t - y_{t-1}$, $d = 2$ gives rise to $\hat{y}_t = (1 - L)^2 y_t = y_t + L^2 y_t - 2L^1 y_t = y_t + y_{t-2} - 2y_{t-1} = (y_{t-2} - y_{t-1}) - (y_{t-1} - y_t)$, etc. Hence, an ARIMA (α, d, β) representation of $\{y_t\}$ is required if $\{y_t\}$ is non-stationary. Equivalently, an ARMA (α, β) model is applied to $\{\hat{y}_t\}$.

Hydrologic time series are generally affected by seasonality. In order to obtain a stationary time series the seasonal component has to be removed. This can be incorporated in the ARIMA framework (e.g., Hillmer and Tiao, 1982).

3.2. ARDL models

Let us consider two time series, $\{x_t\} = (x_1, \ldots, x_n)$ and $\{y_t\} = (y_1, \ldots, y_n)$. The ARDL is a particular type of ARMA models (e.g., De Boef and Keele, 2008; Davidson et al., 1978; Koyck, 1954), in which the MA polynomial is applied to $\{x_t\}$, the time series representing an independent variable (or regressor). Assuming that both time series are stationary, the ARDL model is written as

$$\{\phi \cdot L\} y_t = (\{\theta \cdot L\} - \theta_0 - 1) x_t + \xi_t. \tag{4}$$

This equation can be extended to the case of more than one regressor. In order to obtain a consistent estimation of model coefficients, regressors have not to be cointegrated among themselves (Shin and Pesaran, 1999).

The ARDL model (4) allows one to identify the dynamic relationship between two processes (time series). The influence of the processes involved is distributed over time. If the relationship between y_t and x_t exhibits no delay and persistence, (4) reduces to a generic static model

$$y_t = \theta x_t + \xi_t. \tag{5}$$

3.3. Implementation

The main steps involved in the implementation of an ARMA model are summarized in the following.

(i) Analysis of stationarity. A weak stationarity is generally required for time series, i.e., the mean and auto-covariance have to be constant in time. Several alternative strategies allow one to construct the stationary component of a time series. One way is to apply the differences embedded in the ARIMA framework. An alternative is to employ parametric strategies to model the trend and seasonal components, and then subtract them from the original time series (Box and Jenkins, 1976). We employ the latter strategy by representing time series with an additive model,

$$y_t = q(t) + \sum_{i=1}^{T} v_i \delta\left(i - R\left\{\frac{t}{T}\right\}\right) + \hat{y}_t. \tag{6}$$

Here $\hat{y}_t$ is the stationary counterpart of y_t; $q(t)$ is a polynomial in t, which represents the trend component of y_t; $R\{\cdot\}$ returns the remainder of the division expressed by the argument, $\delta(\cdot)$ is the Dirac delta function, and T is the period associated with the seasonal component. Coefficients of the polynomial $q(t)$ and the coefficients v_i are determined via a multilinear regression.

(ii) Identification of model structure and estimation of model coefficients. The order (α, β) is set according to the following considerations. The auto-correlation function of $\{y_t\}$ provides an indication of the process' persistence that is related to the order α. The parameter β and model coefficients are determined iteratively ($\beta = 1, 2, 3 \ldots$), within a maximum likelihood framework, by relying on a model selection criteria, e.g., AIC (Akaike, 1974)

$$\mathrm{AIC} = \ln\left(\frac{1}{n}\sum_{t=1}^{n} \xi_t^2\right) + \frac{2K}{n}, \tag{7}$$

where K is the total number of estimated coefficients, and n is the size of the time series. According to this criterion, the most favored model, i.e., the value of β in (4), corresponds to the lowest value of AIC.

For a given β, model coefficients are computed via least square estimation in a multiple linear regression. Following Almon (1965), we start by rewriting Eq. (4) in a matrix form, $\mathbf{y} = \mathbf{A} \cdot \mathbf{c} + \boldsymbol{\xi}$, where $\mathbf{y} = (y_1, \ldots, y_n)^\top$; $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$; $\mathbf{c} = (\lambda_0, \lambda_1, \lambda_2, \phi_1, \ldots, \phi_p)^\top$ is the K-dimensional vector with λ_0, λ_1 and λ_2 defined in terms of θ_j as $\theta_j = \lambda_0 + \lambda_1 j + \lambda_2 j^2$; and $\mathbf{A}$ is the $n \times K$ matrix,

$$\mathbf{A} = \begin{pmatrix} z_{0,t_1} & z_{1,t_1} & z_{2,t_1} & y_{t_1-1} & \cdots & y_{t_1-p} \\ z_{0,t_2} & z_{1,t_2} & z_{2,t_2} & y_{t_2-1} & \cdots & y_{t_2-p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ z_{0,t_n} & z_{1,t_n} & z_{2,t_n} & y_{t_n-1} & \cdots & y_{t_n-p} \end{pmatrix} \tag{8}$$

with $z_{0,t} = \sum_{j=0}^{\beta} x_{t-j}$, $z_{1,t} = \sum_{j=0}^{\beta} j x_{t-j}$ and $z_{2,t} = \sum_{j=0}^{\beta} j^2 x_{t-j}$. The vector $\mathbf{c}$ is computed as

$$\mathbf{c} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{y}. \tag{9}$$

This completes the determination of all the coefficients in (4).

(iii) Test on the residuals. The time series $\{\xi_t\}$ has to behave as white noise to satisfy the modeling assumptions. Specifically, the residuals need to be independent, identically distributed random variables, sampled from a normal zero-mean distribution. To verify this condition, we use the Ljung-Box test (Ljung and Box, 1978)

$$H = n(n+2)\sum_{u=1}^{m} \frac{\rho_u^2}{n-u} < \chi_\kappa^2(m - \alpha - \beta), \tag{10}$$

where ρ_u^2 is the autocorrelation coefficient, and m is an integer chosen on the basis of the sample length n. If H fits the χ_κ^2

distribution with significance level $1-\kappa$, then the residual sequence behaves as white noise.

The ARMA models provide a means for both assessment of the underlying mechanisms of a dynamic system and quantitative prediction of its behavior.

4. Results and discussion

Our construction of statistical models of the coastal aquifer relies on time series of top salinity $\{s_t\}$, water table elevation $\{w_t\}$, rainfall $\{r_t\}$, and pumping rate $\{p_t\}$, all measured at the monthly scale. As a preliminary step, each series has to be decomposed into a deterministic trend a stationary component, as required by the ARDL theory (Section 3). This task is accomplished by employing the parametric strategy given by Eq. (6). The resulting zero-mean stationary time series, $\{\hat{s}_t\}$, $\{\hat{w}_t\}$, $\{\hat{r}_t\}$ and $\{\hat{p}_t\}$, are considered in the analysis reported below (hereinafter, for the sake of simplicity, the hats are omitted). Fig. 3 depicts the available time series and their stationary counterparts.

4.1. Statistical models of available data

The data in the coastal strip under consideration mainly come from two piezometers, labeled P_1 and P_2 in Fig. 1. Piezometer P_1 belongs to a network of the University of Bologna (Greggio et al., 2012). It has been used to monitor the phreatic level during several months in 2010 and 2011. Piezometer P_2 is part of the network of the Emilia-Romagna Region Geological Survey. It has been used to sample water levels and salinity, at multiple depths, from 2009 to 2012. The resulting water table elevation (w) and salinity ($\bar{s}$) data (extrapolated at monthly scale) form time series $\{w_{P_1,t}\}$, $\{w_{P_2,t}\}$ and $\{\bar{s}_{P_2,t}\}$, respectively, where P_1 and P_2 refer to the piezometers in which these quantities are measured. The size of time series $\{w_{P_1,t}\}$ is 19 (Fig. 3a), while the size of $\{w_{P_2,t}\}$ and $\{\bar{s}_{P_2,t}\}$ is 36 (Fig. 3b and c, respectively).

Top salinity, i.e., the salinity measured at the top of the phreatic surface, provides an indicator of the impact of groundwater salinization on the ecosystem and biodiversity. The species *Pinus Pinaster*, which is widespread in the study area, thrives if water salinity does not exceed the threshold value of 5–7 g/l (Antonellini and Mollema, 2010). Top salinity is controlled by several factors (e.g., rainfall infiltration, drainage, evapotranspiration from pine trees). Most of these factors, however, are linked to groundwater flow dynamics and directly affect the water table level. Hence, in the following, we investigate how variations in water level at piezometer P_2 affect top salinity. To do this, we assume a lack of delay between the top salinity, $\{s_{P_2,t}\}$, and water table elevation, $\{w_{P_2,t}\}$, time series. This allows us to adopt the static model (5) to represent their relationship

$$s_{P_2,t} = \theta_0 w_{P_2,t} + \xi_t. \tag{11}$$

where $\theta_0 = -16.757$ is derived via least square method.

Many of the key factors affecting top salinity do so indirectly, via the phreatic surface elevation. The first of these factor is precipitation. To investigate the dynamics of natural recharge in the study area we employ an ARDL model. This enables us to account for the time delay, due to the infiltration through the vadose zone, between the rainfall and the water table level. The rainfall data in the area of interest were collected at the daily scale, by a pluviometer whose location is shown in Fig. 1. These daily data are averaged to the monthly scale to form time series $\{r_t\}$ consisting of 48 elements (Fig. 3d); $\{r_t\}$ covers the time period for which measurements of the water table elevation are available.

The water table data collected by piezometers P_1 and P_2 are described by ARDL models

$$w_{P_1,t} = \sum_{i=1}^{\alpha_1} \phi_{1,i} w_{P_1,t-i} + \sum_{j=0}^{\beta_1} \theta_{1,j} r_{t-j} + \xi_t \tag{12}$$

and

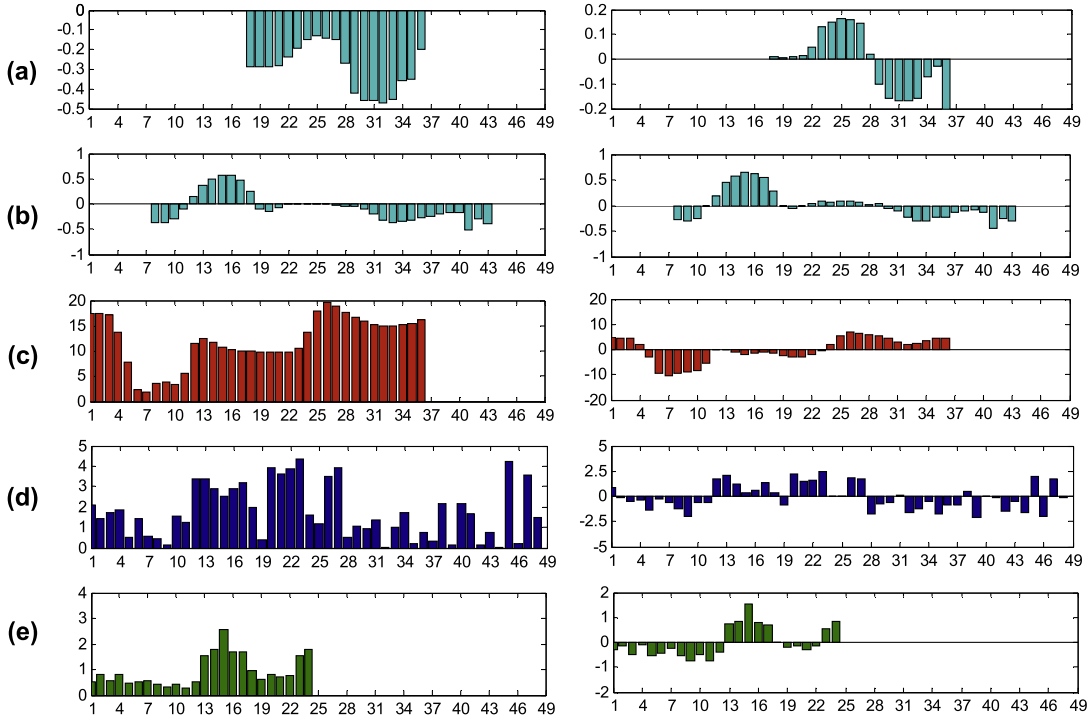

Fig. 3. Monthly time series of (a) $\{w_{P_1,t}\}$ (m asl), (b) $\{w_{P_2,t}\}$ (m asl), (c) $\{s_{P_2,t}\}$ (g/l), (d) $\{r_t\}$ (mm/day), and (e) $\{p_t\}$ (mm/day). Each row comprises an observed time series and its stationary components, from January 2009 forward.

Table 1
Values of the orders and coefficients in the ARDL models (12)–(14).

	α	β	ϕ_1	ϕ_2	ϕ_3	θ_0	θ_1	θ_2
Eq. (12)	2	2	1.1982	−0.5517	–	0.0237	0.0097	0.0018
Eq. (13)	3	2	1.3187	−0.3433	−0.2104	0.0305	−0.0008	−0.0110
Eq. (14)	3	2	0.5981	0.3151	−0.4956	0.1311	0.0728	0.0290

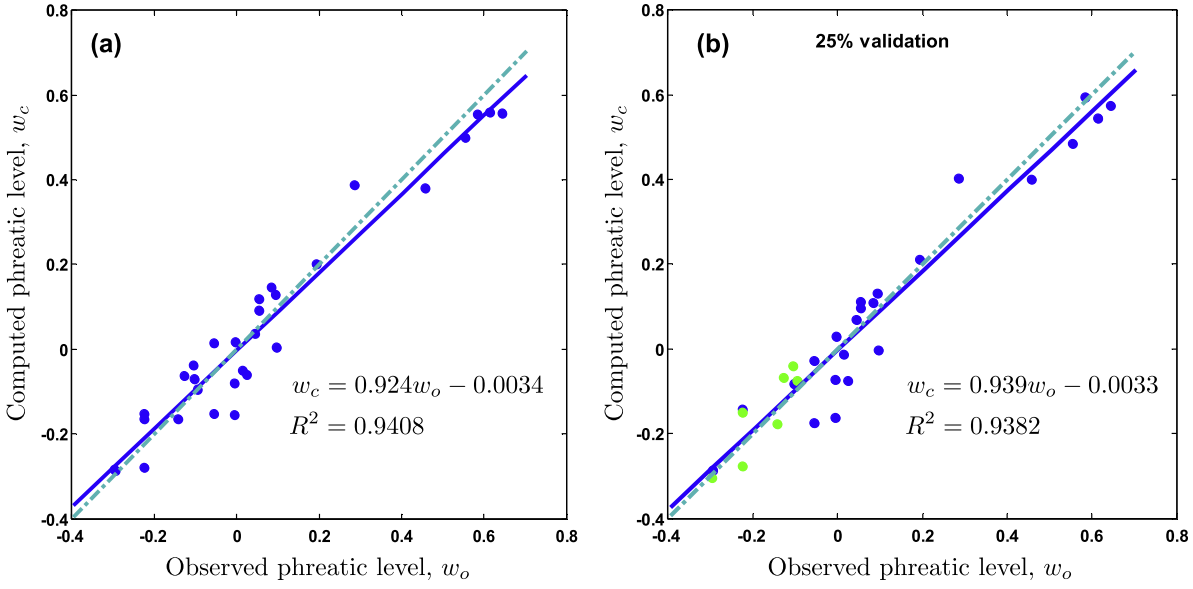

Fig. 4. Phreatic levels observed in piezometer P_2 and predicted with the ARDL model (13) from (a) all the available observations and (b) the first 75% of all the observations (blue), with the remaining data (green) used for validation. The phreatic levels are represented by the stationary components of the time series. The 1:1 straight line corresponds to the perfect agreement between measurements and predictions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

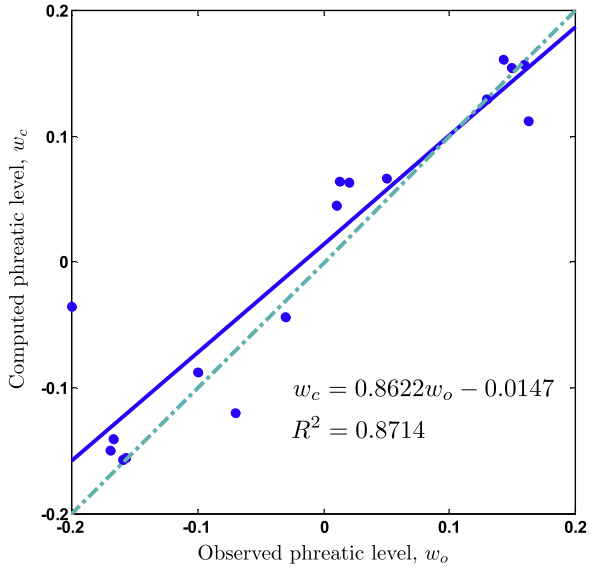

Fig. 5. Phreatic levels observed in piezometer P_1 and predicted with the ARDL model (12). The phreatic levels are represented by the stationary components of the time series. The 1:1 straight line corresponds to the perfect agreement between measurements and predictions.

$$w_{P_2,t} = \sum_{i=1}^{\alpha_2} \phi_{2,i} w_{P_2,t-i} + \sum_{j=0}^{\beta_2} \theta_{2,j} r_{t-j} + \xi_t, \tag{13}$$

respectively. The order parameters α_1 and α_2 in these equations are determined by analyzing the auto-correlation functions of time series $\{w_{P_1,t}\}$ and $\{w_{P_1,t}\}$, respectively. We found the correlation between different elements of time series $\{w_{P_1,t}\}$ to be statistically significant (i.e., to have the autocorrelation coefficient exceeding the 95% confidence interval) if the lag does not exceed 2 months, which yields $\alpha_1 = 2$. A similar analysis of time series $\{w_{P_2,t}\}$ gives rise to $\alpha_2 = 3$. The order parameters β_1 and β_2 are computed together with the remaining coefficients in (12) and (13) combining model selection criteria and multilinear regression following Eqs. (7)–(9). Their values are collected in Table 1.

The phreatic level is also influenced by the drainage system. Water reclamation (pumping) is necessary to allow for agriculture and settlements and is pursued with the aim of maintaining the water level constant over time. The pumping rate time series, $\{p_t\}$, consists of 24 measurements collected at the pumping station (Fig. 3d). A relationship should exist between the pumping rate used in reclamation, $p(t)$, and the rainfall, $r(t)$, since the two have the opposite effect on the water table level. We use an ARDL model,

$$p_t = \sum_{i=1}^{\alpha_3} \phi_{3,i} p_{t-i} + \sum_{j=0}^{\beta_3} \theta_{3,j} r_{t-j} + \xi_t, \tag{14}$$

to establish the relationship between the two time series, $\{p_t\}$ and $\{r_t\}$, constructed from the monthly data of pumping and rainfall collected between January 2009 and December 2010. Following the previously described procedure, we compute the values of the parameters in (13). These are reported in Table 1.

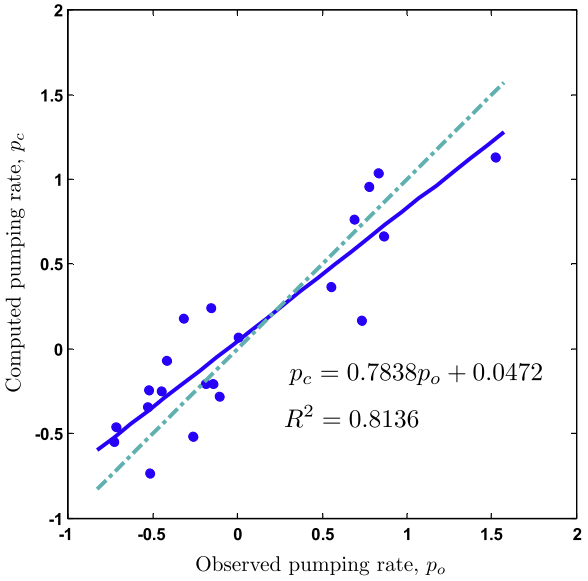

Fig. 6. Pumping rates observed in the field and predicted with the ARDL model (14). The pumping rates are represented by the stationary components of the time series. The 1:1 straight line corresponds to the perfect agreement between measurements and predictions.

4.2. Results

We use the ARDL models (12) and (13) to quantify natural recharge in the area of interest. The order parameter β in these models is related to the time delay between precipitation, $r(t)$, and water table response, $w(t)$. The strategy outlined in Section 3.3 suggests this delay to be about 2–3 months, which is consistent with observations at the site: the phreatic level in autumn is affected by low precipitation occurring during summer, and the effects of heavy winter rainfall are still detected in spring.

The second order parameter, α, represents the persistence (memory) of the phreatic level. The latter is to be expected in a dynamical system, such as a shallow aquifer recharged primarily by precipitation. Our analysis revealed that $\alpha_1 < \alpha_2$ (see Table 1). This is consistent with the fact that piezometer P_1 is located in the vicinity of a river (Fig. 1), which magnifies the dynamic effects of natural recharge, i.e., shortens the system's memory.

In Fig. 4a and b, we compare the phreatic levels observed in piezometer P_2 and predicted with the ARDL model (13). All available observations were used to calibrate the model in Fig. 4a; the model in Fig. 4b used the first 75% of these data for model calibration and the remaining 25% for model validation. In both cases, the points are clustered around the 45° regression line, with virtually negligible spreading. This demonstrates the model's robustness and ability to accurately reproduce the available observations.

A similar analysis is performed for the phreatic levels measured at P_1 and computed with the ARDL model (12). However, given the small size of the time series, all the data are used for model calibration and model validation is omitted. Fig. 5a demonstrates that the accuracy of model's predictions is still acceptable even if decreases in this second case. This is mainly caused by the limited number of available observations; the indirect recharge due to the surface water stream plays a role too.

The results reported in Figs. 4 and 5 imply that temporal patterns of precipitation play an important role in determining the water table levels at the site. This finding is consistent with the earlier observations (Antonellini et al., 2008; Mollema et al.,

Table 2
Values of the angular coefficients, m_{rl}, and coefficients of determination, R^2_{rl}, associated with the regression lines in Figs. 4–6. Autocorrelation coefficients of the residual sequence for the first six lags with the $\mathcal{E} = 95\%$ confidence bounds.

	m_{rl}	R^2_{rl}	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	$\mathcal{E}$
Eq. (12)	0.8622	0.8714	0.035	−0.2849	−0.1341	−0.0286	0.0041	−0.1298	±0.5235
Eq. (13)	0.924	0.9408	0.0039	−0.0481	−0.1162	0.3317	0.0503	0.1653	±0.336
Eq. (14)	0.7838	0.8136	−0.2827	−0.1688	0.0988	−0.1175	0.0063	0.0848	±0.4814

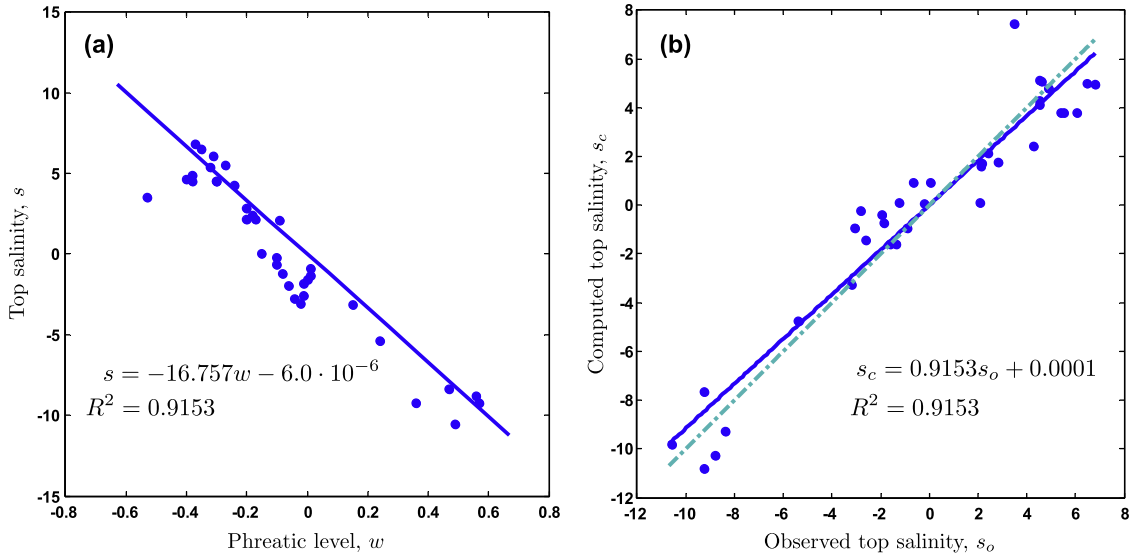

Fig. 7. (a) Relationship between top salinity and phreatic level observed, from the stationary components of the corresponding time series, in piezometer P_2. (b) Top salinity observed in piezometer P_2 and predicted with the static model (11). The salinity values are represented by the stationary components of the time series. The 1:1 straight line corresponds to the perfect agreement between measurements and predictions.

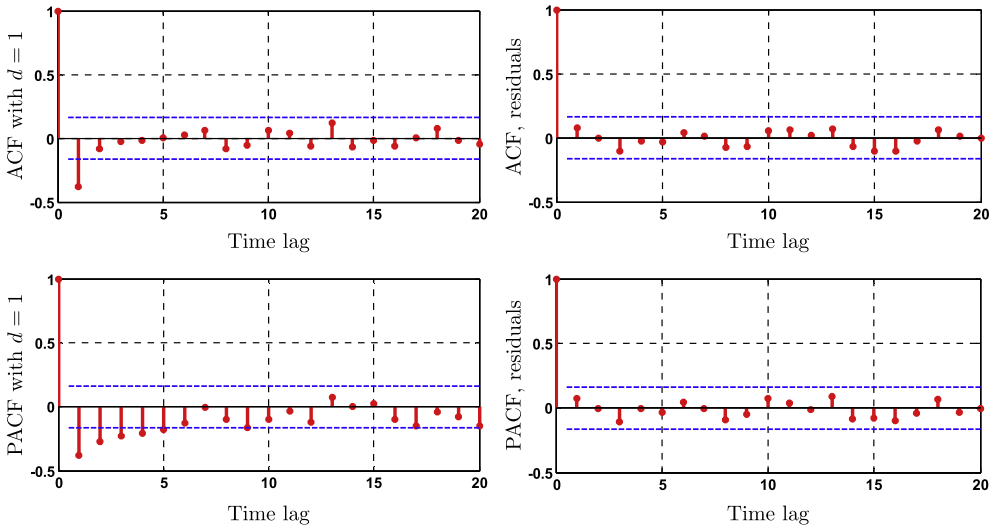

Fig. 8. ACF and PACF of the stationary time series of rainfall (left column) and the residuals associated with the ARIMA (1,1,3) model (right column).

2013) that this shallow coastal aquifer is mainly recharged via rainfall infiltration into the coastal dunes.

The water table in the study area is affected not only by rainfall but also by the land reclamation drainage system, whose purpose is to maintain constant groundwater level. A relationship between the pumping rate associated with land reclamation and the rainfall is provided by the ARDL model (14). Fig. 6 shows that the accuracy of this model is lower then the accuracy of models (12) and (13). This implies that the pumping rate may not be explained by the rainfall alone. This is consistent with the fact that the pump operates almost continuously in time, while the rainfall at the daily scale is intermittent. Furthermore, the drainage channel network is not homogeneous throughout the area, routing the water to the pump with different time delay. Nevertheless, the correlation is sufficiently strong to demonstrate a clear link between the two time series.

Table 2 contains various metrics of the accuracy of the ARDL models (12)–(14). These are angular coefficients (m_{rl}) and coefficients of determination (R^2_{rl}) associated with the regression lines in Figs. 4–6, autocorrelation coefficients (ρ_i) of the residual sequence for the first six lags ($i = 1, \ldots, 6$), and the 95% confidence bounds. We also used (10) to conduct the residuals test, with a significance level of 5%. These results reveal that the residuals behave as white noise, as required for this kind of models; the small size of the available time series gives rise to high sample variances.

Finally, we use the static model (11) to investigate the influence of the phreatic level (independent variable) on the top salinity (dependent variable). Fig. 7a shows the negative correlation between these two variables. Fig. 7b demonstrates the high accuracy of the model's predictions of top salinity, when compared to the observations. The phreatic level acts as an indirect measure of groundwater dilution just below the phreatic surface due to rainfall. Hence, the top salinity is mainly controlled by natural recharge, especially where the sand of the coastal dunes is directly exposed at the surface. In the areas covered by pine forests, groundwater top salinity may also be influenced by the combined effects of water root uptake and evaporation (Mollema et al., 2012). These mechanisms are not included in our analysis.

4.3. Forecasting

Since our models relate top salinity and phreatic levels to rainfall, they may be used to predict effects of changing precipitation

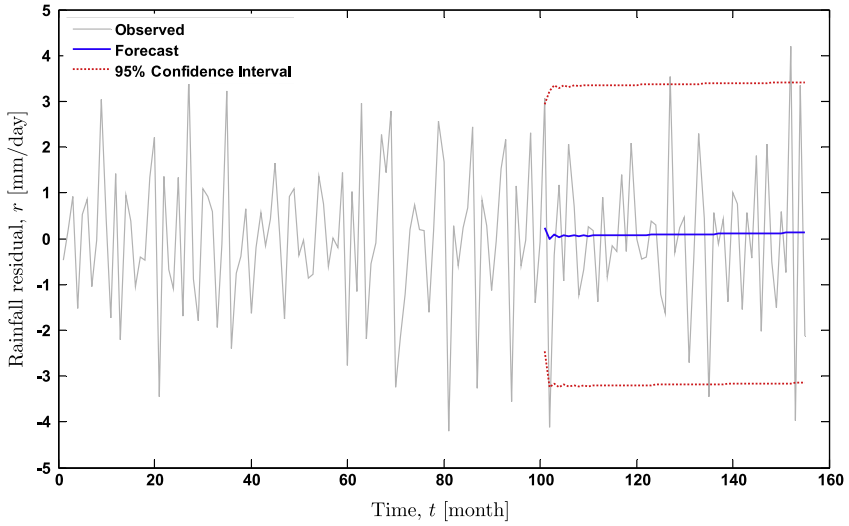

Fig. 9. Forecast of the residuals of the rainfall time series obtained with the ARIMA (1,1,3) model. The predicted stationary time series consists of 55 monthly periods.

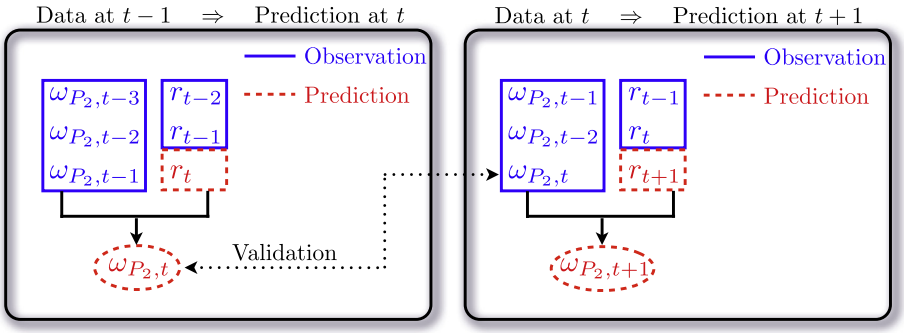

Fig. 10. Work flow for forecast of precipitation and water level and for model validation.

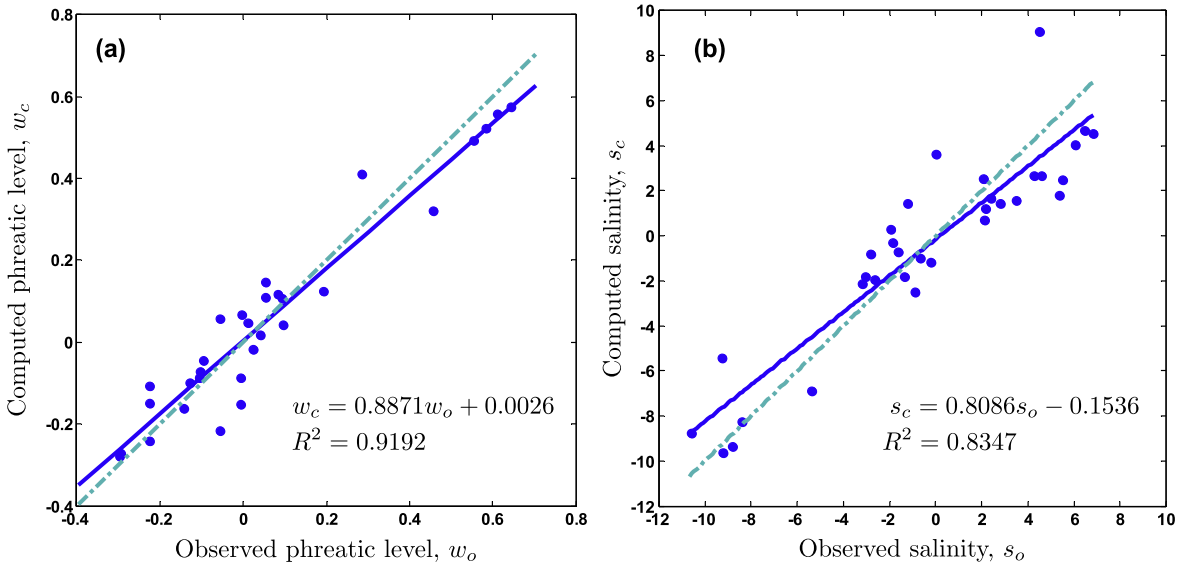

Fig. 11. Observations and predictions of (a) phreatic levels and (b) salinities in piezometer P_2. The data are inferred from the stationary components of the corresponding time series. The predictions rely on the calibration procedure described in Fig. 10. The 1:1 straight line corresponds to the perfect agreement between measurements and predictions.

patterns on the top salinity in the study area. Pluviometric data are typically available over long time periods, which make them suitable for an ARIMA analysis. Consider, for example, the monthly rainfall data recorded from 1990 to 2012. After employing the differentiation (3) with $d = 1$ to render this time series stationary, we compute its autocorrelation function (ACF) and partial correlation function (PACF). These are shown in the left column of Fig. 8, which also reveals the absence of a seasonal component in this time series. The ACF cuts off sharply after the first lag, while PACF decays more gradually, necessitating the use of the moving average (Box and Jenkins, 1976). Among different combinations of the order parameters, an ARIMA (1,1,3) model yields the smallest variance. Its model coefficients are $\phi_1 = -0.8393$, $\theta_1 = -0.8607$, $\theta_2 = -0.8253$, and $\theta_3 = 0.6925$. The ACF and PACF of the residual sequence are depicted in the right column of Fig. 8. The model prediction of the last 35% of available data, which were used for validation, is shown in Fig. 9 together with the 95% confidence interval.

Finally, we use the ARIMA model to forecast the precipitation pattern in the area of interest. Since its predictive power deteriorates with the forecasting time horizon, the model has to be updated (recalibrated) as new data become available. Fig. 10 depicts a recalibration procedure, which consists of the following steps. One future value of rainfall is predicted at each time step with the ARIMA model. Then, the value of water level, at the same time, is computed by means of the ARDL model, as described in the previous section. When the observation corresponding to the predicted value of water level becomes available, the two are compared for validation purposes. Stationary components of the top salinity and water level observed at well P_2 are compared with their predicted counterparts in Fig. 11. The water levels were predicted with our recalibration procedure, and the corresponding values of top salinity were obtained by means of the generic static model. This comparison serves to validate the proposed data-driven models. In general, time horizon, over which predictions remain accurate, increases with the time series' length.

5. Summary

Our results provide an insight into the underlying mechanisms of groundwater salinization in reclaimed low coastal plains and can serve as predictive models. We focused on a specific study area in Ravenna (Italy), in which groundwater salinity is critical for coastal vegetation. Specifically, we considered a coastal strip, which provides a habitat for a valuable pine forest. Pine trees, a species not typical of this environment, require both a deep water table and a limited groundwater salinity to thrive. This places a

premium on accurate forecasting and control of groundwater levels.

We employed top salinity as a key metric of groundwater salinization and established a functional dependence between it and water table level, rainfall, and reclamation pumping. Our modeling strategy relies on the auto-regressive moving average (ARMA) framework. The resulting predictive models were validated by comparison with data. We demonstrated that data-driven approaches may provide useful information in situations where physics-based models have only limited success in characterizing the phenomenon of interest. Moreover, these approaches can be used to assist in building physics-based models through a preliminary interpretation of available observations.

Both ARMA and ARDL models are predicated on the assumption that time series under investigation are stationary or become so after de-trending, i.e., decomposing a time series into its (possibly time-dependent) mean and zero-mean fluctuations about the mean (the so-called residuals). Implicit in the ARMA and ARDL models with de-trending is the notion that the residuals contain sufficient informations to establish the interdependencies between the underlying time series and that the average components of these time series have the same relations as the residuals do.

An essentially nonstationary time series is a time series whose residuals remain nonstationary after de-trending, e.g., it might have a time-dependent variance. While important in many applications, e.g., forecasting an aquifer's response to long-term climate change, analysis of essentially nonstationary processes lies outside the scope of the present study.

Acknowledgements

This work was supported in part by Università di Bologna RFO (Ricerca Fondamentale Orientata) 2012 and 2013, by the Air Force Office of Scientific Research under Grant No. FA9550-12-1-0185, and by the National Science Foundation under Grant No. EAR-1246315.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723.
Almon, S., 1965. The distributed lag between capital appropriations and expenditures. Econometrica 33, 178–196.
Amorosi, A., Colalongo, M., Fiorini, F., Fusco, F., Pasini, G., Vaiani, S., Sarti, G., 2004. Palaeogeographic and palaeoclimatic evolution of the Po Plain from 150-ky core records. Global Planet. Change 40, 55–78.
Amorosi, A., Centineo, M.C., Colalongo, M.L., Fiorini, F., 2005. Millennial-scale depositional cycles from the Holocene of the Po Plain, Italy. Mar. Geol. 222–223, 7–18.
Antonellini, M., Mollema, P.N., 2010. Impact of groundwater salinity on vegetation species richness in the coastal pine forests and wetlands of Ravenna, Italy. Ecol. Eng. 36, 1201–1211.
Antonellini, M., Mollema, P., Giambastiani, B., Bishop, K., Caruso, L., Minchio, A., Pellegrini, L., Sabia, M., Ulazzi, E., Gabbianelli, G., 2008. Salt water intrusion in the coastal aquifer of the southern Po Plain, Italy. Hydrogeol. J. 16, 1541–1556.
Antonellini, M., Allen, D.M., Mollema, P.N., Capo, D., Greggio, N., 2015. Groundwater freshening following coastal progradation and land reclamation of the Po Plain, Italy. Hydrogeol. J. 23 (5), 1009–1026.
Armaroli, C., Ciavola, P., Perini, L., Calabrese, L., Lorito, S., Valentini, A., Masina, M., 2012. Critical storm thresholds for significant morphological changes and damage along the Emilia-Romagna coastline, Italy. Geomorphology 143–144, 34–51.
Barbier, E.B., Hacker, S.D., Kennedy, C., Koch, E.W., Stier, A.C., Silliman, B.R., 2011. The value of estuarine and coastal ecosystem services. Ecol. Monogr. 81, 169–193.
Baú, D., Gambolati, G., Teatini, P., 2000. Residual land subsidence near abandoned gas fields raises concern over northern Adriatic coastland. EOS, Trans. Am. Geophys. Union 81, 245–249.
Bear, J., Cheng, A.H.D., Sorek, S., Ouazar, D., Herrera, I., 1999. Seawater Intrusion in Coastal Aquifers: Concepts, Methods and Practices, vol. 14. Springer, The Netherlands.
Beck, N., Katz, J.N., 2011. Modeling dynamics in time-series-cross-section political economy data. Annu. Rev. Polit. Sci. 14, 331–352.
Bondesan, M., Castiglioni, G.B., Elmi, C., Gabbianelli, G., Marocco, R., Pirazzoli, P.A., Tomasin, A., 1995. Coastal areas at risk from storm surges and sea level rise in north-eastern Italy. J. Coastal Res. 11, 1354–1379.
Box, G.E.P., Jenkins, G., 1976. Time Series Analysis: Forecasting and Control. Holden-Day Series in Time Series Analysis and Digital Processing, second ed. Holden-Day, San Francisco, CA.
Cheng, A.H.D., Konikow, L.F., Ouazar, D., 2001. Special issue of transport in porous media on 'seawater intrusion in coastal aquifers'. Transport Porous Med. 43, 1–2.
Ciabatti, M., 1968. Gli antichi delta del po anteriori al 1600. In: Atti del convegno internazionale di studi sulle antichità di Classe, Ravenna, pp. 23–33.
Custodio, E., 2010. Coastal aquifers of Europe: an overview. Hydrogeol. J. 18, 269–280.
Davidson, J.H., Hendry, D.F., Srba, F., Yeo, S., 1978. Econometric modelling of the aggregate time-series relationship between consumers' expenditures and income in the United Kingdom. Economic J. 88, 61–92.
De Boef, S., Keele, L., 2008. Taking time seriously: dynamic regression. Am. J. Polit. Sci. 52, 184–200.
De Louw, P.G.B., Oude Essink, G.H.P., Stuyfzand, P.J., van der Zee, S.E.A.T.M., 2010. Upward groundwater flow in boils as the dominant mechanism of salinization in deep polders, The Netherlands. J. Hydrol. 394, 494–506.
Feagin, R.A., Sherman, D.J., Grant, W.E., 2005. Coastal erosion, global sea-level rise, and the loss of sand dune plant habitats. Front. Ecol. Environ. 3, 359–364.
Gambolati, G., Teatini, P., Gonella, M., 2002. GIS simulations of the inundation risk in the coastal lowlands of the Northern Adriatic Sea. Math. Comput. Model. 35, 963–972.
Giambastiani, B.M.S., Antonellini, M., Oude Essink, G.H.P., Stuurman, R.J., 2007. Saltwater intrusion in the unconfined coastal aquifer of Ravenna (Italy): a numerical model. J. Hydrol. 340, 91–104.
Giambastiani, B.M.S., Greggio, N., Pacella, K., Iodice, A., Antonellini, M., 2014. Effect of forest fire on coastal aquifer salinization and freshwater availability. In: 23rd Salt Water Intrusion Meeting, Husum, Germany, pp. 125–128.
Greggio, N., Mollema, P., Antonellini, M., Gabbianelli, G., 2012. Irrigation management in coastal zones to prevent soil and groundwater salinization. In: Abrol, V., Sharma, P. (Eds.), Resource Management for Sustainable Agriculture. InTech, pp. 21–48 (Chapter 2).
Grootjans, A.P., Ernst, W.H.O., Stuyfzand, P.J., 1998. European dune slacks: strong interactions of biology, pedogenesis and hydrology. Trends Ecol. Evol. 13, 96–100.
Hamilton, J., 1994. Time Series Analysis. Princeton University Press.
Hillmer, S.C., Tiao, G.C., 1982. An ARIMA-model-based approach to seasonal adjustment. J. Am. Stat. Assoc. 77, 63–70.
Koyck, L.M., 1954. Distributed lags and investment analysis. Contributions to Economic Analysis, vol. 4. North-Holland, Amsterdam, The Netherlands.
Langevin, C.D., Thorne, D., Jr. Dausman, A., Sukop, M., Guo, W., 2007. SEAWAT Version 4: a computer program for simulation of multi-species solute and heat transport. In: U.S. Geological Survey Techniques and Methods Book, U.S. Geological Survey, vol. 6, p. 39 (Chapter A22).
Ljung, G.M., Box, G.E.P., 1978. On a measure of lack of fit in time series models. Biometrika 65, 297–303.
Marchesini, L., Amorosi, A., Cibin, U., Zuffa, G., Spadafora, E., Preti, D., 2000. Sand composition and sedimentary evolution of a late Quaternary depositional sequence, Northwestern Adriatic Coast, Italy. J. Sediment. Res. 70, 829–838.
Marconi, V., Antonellini, M., Balugani, E., Dinelli, E., 2011. Hydrogeochemical characterization of small coastal wetlands and forests in the Southern Po plain (Northern Italy). Ecohydrology 4, 597–607.
McGranahan, G., Balk, D., Anderson, B., 2007. The rising tide: assessing the risks of climate change and human settlements in low elevation coastal zones. Environ. Urban. 19, 17–37.
Mollema, P., Antonellini, M., Gabbianelli, G., Laghi, M., Marconi, V., Minchio, A., 2012. Climate and water budget change of a Mediterranean coastal watershed, Ravenna, Italy. Environ. Earth Sci. 65, 257–276.
Mollema, P., Antonellini, M., Dinelli, E., Gabbianelli, G., Greggio, N., Stuyfzand, P., 2013. Hydrochemical and physical processes influencing salinization and freshening in mediterranean low-lying coastal environments. Appl. Geochem. 34, 207–221.
Oude Essink, G.H.P., Kooi, H., 2012. Land-subsidence and sea-level rise threaten fresh water resources in the coastal groundwater system of the Rijnland water board, The Netherlands. In: Treidel, H., Martin-Bordes, J.L., Gurdak, J.J. (Eds.), Climate Change Effects on Groundwater Resources: A Global Synthesis of Findings and Recommendations. CRC Press, pp. 227–248, Chapter 1.
Oude Essink, G.H.P., van Baaren, E.S., de Louw, P.G.B., 2010. Effects of climate change on coastal groundwater systems: a modeling study in the Netherlands. Water Resour. Res. 46, 1–16.
Pitman, M., Läuchli, A., 2002. Global impact of salinity and agricultural ecosystems. In: Läuchli, A., Lüttge, U. (Eds.), Salinity: Environment-Plants-Molecules. Springer, The Netherlands, pp. 3–20.
Post, V.E.A., der Plicht, H.V., Meijer, H.A.J., 2003. The origin of brackish and saline groundwater in the coastal area of the Netherlands. Neth. J. Geosci. 82, 133–147.
Shahin, M., Van Oorschot, H., De Lange, S., 1993. Statistical Analysis in Water Resources Engineering. A.A. Balkema, Rotterdam, The Netherlands.
Shin, Y., Pesaran, M., 1999. An autoregressive distributed lag modelling approach to cointegration analysis. In: Strom, S. (Ed.), Econometrics and Economic Theory in the 20th Century: The Ragnar Frish Centennial Symposium. Cambridge University Press, pp. 371–413.

Shumway, R.H., Stoffe, D.S., 2005. Time Series Analysis and Its Applications. Springer-Verlag, Secaucus, NJ.
Small, C., Nicholls, R., 2003. A global analysis of human settlement in coastal zones. J. Coastal Res. 19, 584–599.
Stefani, M., Vincenzi, S., 2005. The interplay of eustasy, climate and human activity in the late Quaternary depositional evolution and sedimentary architecture of the Po Delta system. Mar. Geol. 222–223, 19–48.
Teatini, P., Ferronato, M., Gambolati, G., Gonella, M., 2006. Groundwater pumping and land subsidence in the Emilia-Romagna coastland, Italy: modeling the past occurrence and the future trend. Water Resour. Res. 42, 1–19.
Ulazzi, E., Antonellini, M., Gabbianelli, G., 2008. Saltwater intrusion in a unconfined coastal aquifer: the case study of Cervia (North Adriatic sea, Italy). In: Meire, P., Coenen, M., Lombardo, C., Robba, M., Sacile, R. (Eds.), Integrated Water Management: Practical Experiences and Case Studies, Nato Science Series IV: Earth and Environmental Sciences, vol. 80. Springer, The Netherlands, pp. 295–308.
Veggiani, A., 1974. Le ultime vicende geologiche del ravennate. In: Vitale di Ravenna, S. (Ed.), Influenza di insediamenti industriali sul circostante ambiente naturale – Studio sulla pineta di. Compositori, Bologna, pp. 48–58.
Wang, H.R., Wang, C., Lin, X., Kang, J., 2014. An improved ARIMA model for precipitation simulations. Nonlinear Process. Geophys. 21, 1159–1168.