Noname manuscript No.
(will be inserted by the editor)

Predicting seasonal and hydro-meteorological impact in environmental variables modelling via Kalman filtering

A. Manuela Gonçalves · Marco Costa

Received: date / Accepted: date

Abstract This study focuses on the potential improvement of environmental variables modelling by using linear state-space models, as an improvement of the linear regression model, and by incorporating a constructed hydro-meteorological covariate. The Kalman filter predictors allow to obtain accurate predictions of calibration factors for both seasonal and hydro-meteorological components. This methodology can be used to analyze the water quality behaviour by minimizing the effect of the hydrological conditions. This idea is illustrated based on a rather extended data set relative to the River Ave basin (Portugal) that consists mainly of monthly measurements of dissolved oxygen concentration (DO) in a network of water quality monitoring sites. The hydro-meteorological factor is constructed for each monitoring site based on monthly precipitation estimates obtained by means of a rain gauge network associated with stochastic interpolation (Kriging). A linear state-space model is fitted for each homogeneous group (obtained by clustering techniques) of water monitoring sites. The adjustment of linear state-space models is performed by using distribution-free estimators developed in a separate section.

Keywords Hydrological basin · Water quality · State-space modelling · Kalman filter · Distribution-free estimation

A. Manuela Gonçalves
Departamento de Matemática e Aplicações, Universidade do Minho
Campus de Azurém da Universidade do Minho, 4800-058 Guimarães, Portugal
CMAT - Centro de Matemática da Universidade do Minho
E-mail: mneves@math.uminho.pt

Marco Costa
Escola Superior de Tecnologia e Gestão de Águeda, Universidade de Aveiro
Apartado 473, 3750-127 Águeda, Portugal
CMAF - Centro de Matemática e Aplicações Fundamentais da Universidade de Lisboa
E-mail: marco@ua.pt

1 Introduction

The administration of hydrologic resources has been deserving a special prominence in the context of domestic and international politics in order to solve the complexity and the uncertainty of the problems associated with a world-wide and local scale of sustainable administration (environmental, social, and economical) of natural water resources.

The river basin, which is the primordial unity of water resources planning and management, is usually submitted to pressures and changes due to human activities. At a river basin scale there is a need to establish a methodology for systematic data monitoring, for the characterization of surface water quality and for the correct analysis of collected data (Vega et al 1998). Surface water quality monitoring has as its main objective the characterization of water resources, as well as the monitoring of its space-time evolution in order to achieve an appropriate administration.

A river is a system comprising both the main course and its tributaries, carrying the one-way flow of a significant load of matter in dissolved and particulate phases from both natural and anthropogenic sources (Shrestha and Kazama 2007). This study focuses on a rather extended data set relative to the River Ave's basin in Northwest Portugal and consists mainly of monthly measurements of physical-chemical and microbiological variables in a network of water quality monitoring sites and of monthly precipitation in a rain gauge network of meteorological monitoring sites. The River Ave's hydrological basin has an approximate area of 1400 Km^2 (from its source in Serra da Cabreira to its mouth in Vila do Conde), it's 101 Km in length and its average flow at the mouth is of about 40 m^3/s. Its main adjacent streams are the River Este (flowing from the North) and the Rivers Selho and Vizela (from the South). In the last thirty years, the River Ave's hydrological basin, with the exception of its upstream areas, has been subjected to a growing rhythm of untreated effluents discharges from industrial activities, namely from the textile sector strongly implanted in this region. All this situation is instrumental for the water quality deterioration, resulting in inappropriate water for several uses: human consumption, industrial use, recreational uses, fishing and irrigation, thus posing a serious danger for public health (Oliveira et al 2005). The River Ave differs from the other Northern region rivers not only because of its high pollution levels but also due to the large space-time variability of pollutants concentration. The water quality measurements failed to comply with the objectives of minimum quality for surface waters prescribed by the Portuguese legislation. The Central Administration, through the Regional Directory for the Northern Environment and Natural Resources (DRAN) and the Institute of Water (INAG) monthly monitored the surface water quality along the River Ave and its main streams since 1988 by means of a monitoring net encompassing 20 water monitoring sites that in 1998 was redimensioned in order to comply with the new legislation. This network has been constantly restructured since 2007, in order to implement its chemical status monitoring (2007) and, more recently, its ecological status monitoring (2009), as stipulated by the Water Framework Directive (Machado et al 2010).

Multivariate statistical analysis has been widely applied in water quality assessment and sources apportionment of water over the last years (Wunderlin et al. 2001; Simeonov et al. 2003; Shrestha and Kazama 2007). In several works, multivariate statistical analyses are applied to sets of water quality variables, usually quantitative analytical data consisting of physico-chemical variables. If the goal is to investigate water quality evaluation in its time-space variations as in Helena et al. (2000), or the natural and anthropogenic origins of contaminants in surface or ground water as in Ato et al. (2010), the most suitable and applied approach is the principal components analysis (Liu et al. 2003; Lischeid 2009; Varol and Sen 2009). In some practical studies, there is data available from a group of sample sites, usually water monitoring sites, which is useful to perform several statistical methodologies: for instance, correlation analysis parametric and non-parametric tests (Elhatip et al. 2008).

When a predict model is needed, the linear regression has been the most applied approach (e.g. Gonçalves and Alpuim 2011; Renwick et al. 2009). However, statistical models with fixed effects are unlikely to yield a good predictive accuracy, particularly in situations where the predictor and predictand relationship changes over time (Kokic 2010). This issue has been previously acknowledged in environmental data: Costa and Alpuim (2011) consider state-space models in the calibration of radar precipitation measures and Charles et al. (2004) and Greene et al. (2008) have taken hidden Markov Chain models to represent an evolving climate system in statistical downscaling. Costa and Gonçalves (2011) proposed a methodology which combines the analysis of a set of sample sites–which were obtained by means of clustering procedures–with the adjustment of predict regression models and state-space models, in particular considering trends and seasonal components. However, it was demonstrated that state-space models improved the predictions accuracy in comparison with the linear regression models.

In this study, a linear state-space model is proposed for modelling continuous physical and chemical monitoring data. The model was applied to dissolved oxygen concentrations levels (DO) (mg/l) in 8 monitoring sites in the River Ave's basin over a 12-year period (1998-2009). Adequate dissolved oxygen is necessary for good water quality and it is one of the most important variables in the assessment of river water quality and pollution grade.

The proposed methodology starts by using a multivariate statistical approach–cluster analysis–to classify the water quality monitoring sites into homogeneous space-time groups based on the DO quality variable which was selected and considered relevant to characterize the water quality. In a recent work, Costa and Gonçalves (2011) show that a set of water quality monitoring sites can be modelled by applying cluster techniques that minimize the number of models.

One of the problems faced by meteorologists and hydrologists that study spatial rainfall patterns is the interpolation of data from irregularly spaced rain gauges in order to determine mean area rainfalls or to characterize rainfall variability within a region or catchment (Dirks et al 1998; Ciach and Krajewski 2006). Many hydrological and ecological studies recognize the importance of characterizing the time-space variability of precipitation in a geographical area (Goodrich et al 1995), for it is essential to estimate the hydrological balance. Water quality in a given location is the reflex of the dominant conditions in the source basin of that location, namely the hydro-meteorological factors. The behaviour of the space-time quality variable is associated with the flow variation (variable dilution effect), which in turn is generally related to the seasonal rainfall variation.

We present the problem of area precipitation measurement in order to estimate a hydro-meteorological factor that will be used in the modelling of the surface water quality of river basins, particularly for the dissolved oxygen variable. A hydro-meteorological factor is constructed for each quality monitoring site (totalling 8 sites) based on the analysis of the space-time behaviour of the precipitation (monthly total) observed in a rain gauge network constituted by a total of 19 meteorological sites located in the area of the River Ave's basin, between 1931-2009. A geostatistical approach and ordinary Kriging method was chosen with the main goal of identifying models which estimate monthly average rainfall in a sub-basin associated with a water quality monitoring site where there are no observed values. Through stochastic interpolation (Kriging) it is estimated the mean area rainfall during each month in the area of influence of each water quality monitoring site: this covariate will integrate a hydro-meteorological component that is crucial in any water quality modelling process.

Finally, for each cluster, a linear state-space model was fitted to modelling the DO concentration quality variable by taking into account the seasonal variation throughout the year and the estimated hydro-meteorological factor. The results demonstrate the effectiveness and advantages of modelling water quality variables according to this approach, allowing to identify two different components as a seasonal and a hydro-metereological factor.

2 Data set description

Northern Environment and Natural Resources (DRAN) and the Institute of Water (INAG) monthly monitor surface water quality along the River Ave and its main adjacent streams with a net of monitoring sites that comprises more than 23 variables to assess river water quality: industry, domestic wastewater, agriculture, wastewater treatment plants. In total, eight water monitoring sites are considered in this study: five located in the River Ave's mainstream–*Cantelães* (CANT), *Taipas* (TAI), *Riba d'Ave* (RAV), *Santo Tirso* (STI), and *Ponte Trofa* (PTR)–and *Golães* (GOL), *Ferro* (FER), and *Vizela Santo Adrião* (VSA) in the adjacent stream River Vizela. These eight monitoring sites result from the restructuring of the water quality monitoring network in 1998, which implied the closure of other previous sites, and so the data set reports to the period between May 1998 and December 2009. Table 1 summa-

Table 1 Minimum, maximum, mean, standard deviation and missing data rate of water quality variable DO concentration at the 8 monitoring sites in the River Ave's basin

Monitoring Site	CANT	TAI	RAV	STI	PTR	GOL	FER	VSA
Minimum	7.4	6.6	1.8	1.7	2.4	7.3	7.3	7.2
Maximum	12.8	11.72	11.7	12.0	11.7	11.7	11.7	12.4
Mean	9.86	9.32	8.40	8.13	7.94	9.58	9.59	9.67
Standard deviation	1.06	1.13	1.82	2.16	1.92	1.05	1.08	1.13
Missing data rate	7.1%	8.6%	0.7%	1.4%	2.1%	8.6%	5.0%	8.6%

rizes basic statistics for the monthly measurements of the DO water quality variable at the 8 monitoring sites during the above-mentioned period.

DO concentration is an important indicator since most aquatic fauna and flora need oxygen to survive. The river system both produces and consumes oxygen. If more oxygen is consumed than it is produced, dissolved oxygen levels decline and some sensitive animals and plants could disappear. DO is measured in milligrams per liter. Milligrams per liter (mg/l) is the amount of oxygen in a liter of water and it is the same as "parts per million" or *ppm*. Dissolved oxygen concentration is probably the most important factor in assessing the health of a water body, but other factors outside the water managers direct control also determine a water body's health to a variable extent. Organic pollution is the most common type of pollution in this basin and, consequently, a frequent problem is a deficit of DO concentration. This problem is aggravated by the existence of a sequence of small dams in the River Ave and in its main adjacent rivers (Costa and Gonçalves 2011).

3 Cluster analysis

Taking into account previous works based on hydrological river basins (Shresta and Kazama 2007; Costa and Gonçalves 2011), a cluster analysis (CA) was performed for grouping monitoring sites with similar water quality characteristics in time, based on the DO concentration levels. Furthermore, this type of analysis allows reducing the number of models in the modelling process.

CA is a group of multivariate techniques whose primary purpose is to assemble objects based on their characteristics. Hierarchical agglomerative clustering is the most common approach, providing intuitive similarity relationships between any given sample and the entire data set, and is typically illustrated by a dendrogram (McKenna 2003).

In this study, hierarchical agglomerative CA was performed on the raw data set by means of Ward's method. Ward's method uses a variance approach to evaluate the distances between clusters, in an attempt to minimize the sum of squares (SS) of any two clusters that can be formed at each step. As these types of algorithms operate on dissimilarities, our first task is to build a dissimilarity matrix based on some measure of dissimilarity that can be applied to any two

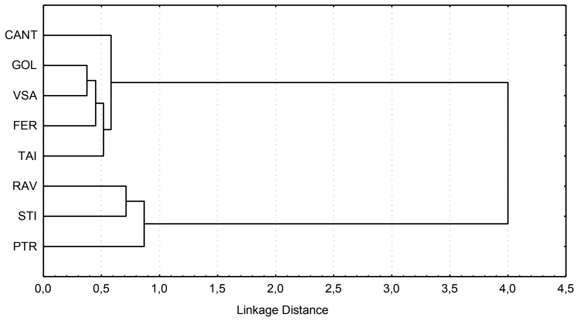

Fig. 1 Dendrogram showing clustering monitoring sites according to DO characteristics based on Ward's method

monitoring sites i and j. In this case, the main problem is that, for all locations, there are no observations for all months under study. Therefore, let us consider x_{it} the value of the DO quality variable measured at location i in time t. The Euclidean distance at this time instant between sites i and j is given by the expression

$$dist_{ij}(t) = |x_{it} - x_{jt}| \, .$$

We use the dissimilarity measure that corresponds to the average of this distance over all months t where there is observed value of the DO quality variable with measurements in the two sites, i.e.

$$d_{ij} = \frac{1}{\# M_{ij}} \sum_{t \in M_{ij}} |x_{it} - x_{jt}| \, , \quad i, j = 1, \ldots, 8,$$

where M_{ij} is the set of all months with DO measured in both sites i and j. Hence, this dissimilarity measure is simply a variant of the average Euclidian distance adjusted to our situation, where the number of sampled sites differs on a monthly basis. This methodology was based on the previous work of Gonçalves and Alpuim (2011).

The monitoring sites dendrogram obtained by means of Ward's method is shown in Figure 1. It has a cophenetic correlation coefficient of 0.85 (i.e., the correlation between the actual dissimilarities as recorded in the original dissimilarity matrix, and the dissimilarities which can be found in the dendrogram).

Two well-differentiated clusters were observed and the results confirm previous knowledge about the effluents discharge according to the economic activities located along the River Ave's basin. Also, the effects of these discharges in water quality vary according to natural and geographical/economical reasons. Cluster I is composed by monitoring sites CANT, TAI, GOL, FER, and VSA. There is a set of locations which have the best water quality indicators (the highest values obtained from the DO concentration), including sites situated

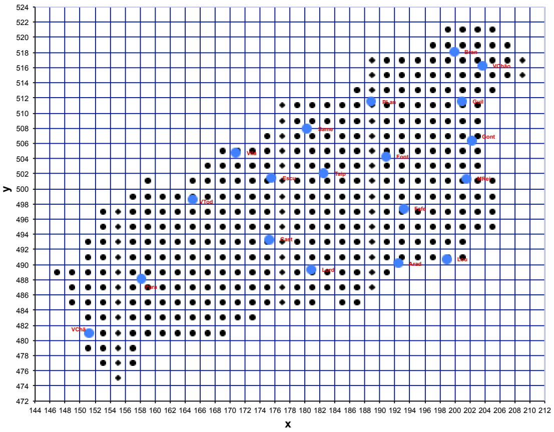

Fig. 2 Spatial distribution of 19 meteorological monitoring sites in the River Ave basin and its discretization in 368 points

upstream the Rivers Ave and Vizela (CANT corresponds to the source of River Ave); these monitoring sites receive pollution mostly from domestic wastewater and from agricultural and manure discharges. In Cluster II, comprised of the three monitoring sites RAV, STI, and PTR located in the River Ave near the most polluted area of the Ponte Trofa and Santo Tirso Municipalities, there is a growing urban population and a high concentration of industrial activity, and it is also where the Ave receives similarly polluted waters from its adjacent rivers (Selho and Vizela), and, consequently, these sites present the worst water quality.

The results of CA confirm the expected behaviour of space-time dynamics of DO concentration observed in the 8 monitoring sites.

4 The hydro-meteorological factor

A hydro-meteorological factor is constructed and will be used as a covariate in the modelling process. This covariate will integrate a hydro-meteorological component that is recognized as crucial in any water quality modelling process. One covariate is constructed for each water monitoring site based on the estimate of the monthly mean precipitation of its influence region.

Figure 2 shows the region corresponding to the River Ave's hydrological basin (approximately 1400 Km^2), which is discretized in 368 points (each point corresponding to cell centers of 2 Km x 2 Km), and the 19 meteorological stations located in this hydrological basin.

The data is available under Cartesian (X, Y) co-ordinates (X - distance to meridian (km) and Y- distance to the perpendicular (km)) coincident with those of the military maps under a 1/25000 scale. The River Ave's hydrological basin is situated between the co-ordinates $X = 147$ Km to East, $X = 209$ Km

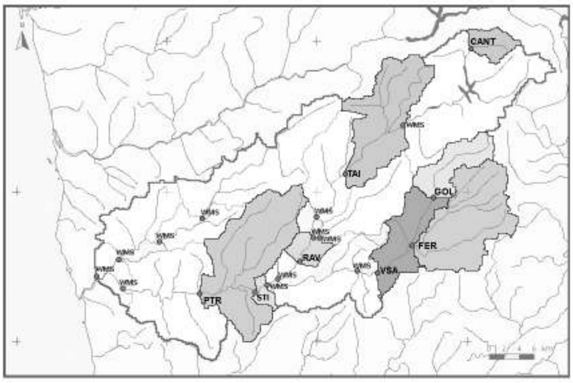

Fig. 3 The limits of the hydrological basins as defined by the water quality monitoring sites (WMS)

to West, $Y = 475$ Km to South and $Y = 521$ Km to North. The adopted measuring unit is the millimeter (mm). The average density of each monitoring site is of about one site for every 73 Km^2, which the World Meteorological Organization considers enough to rigorously define the space-time rainfall variability in this kind of temperate-climate region, and to define the average precipitation values occurred in the basin. In the Northwest of Portugal summer is dry and winter is mild with plenty of rain. So, the highest levels of precipitation take place between October and March: this represents 75% of the yearly precipitation.

Firstly, for each water monitoring site, the monthly mean area precipitation was computed in its influence region based on the average point prediction. Our main goal is to identify models which estimate monthly average rainfall in places where there are no observed values (in the monitoring points of water quality), with the help of the rainfall spatial distribution in other locations, from a set of rain gauges. Due to the large rainfall space-time variability, the precise evaluation, in real time, of mean area estimates poses a difficult problem. To accomplish this, we propose a Kriging stochastic methodology. In general, geostatistical methods are statistical techniques developed to interpolate spatially autocorrelated variables, where the spatial coordinates may either identify geographical location or a position in a generic two-dimensional space. Goovaerts (2000) compared seven techniques used to map monthly data for the Algarve region in Southern Portugal and concluded that geostatistical Kriging methods are better than traditional simple techniques (Thiessen, inverse square distance, regression). In particular, Kriging is a method for optimizing the estimation, under certain conditions, of a quantity that is distributed in space and measured at a network of points (Journel and Huijbregts 1978; De Marsily 1986; Isaaks and Srivastava 1989; Rossi et al. 1992; Chokmain and Ouarda 2004).

Figure 3 shows the River Ave's hydrological basin with the influence areas delineated, and this methodology overlaps the areas of the 20 hydrological sub-basins (linked to the 20 water quality sampling sites). In this context, the influence regions of each water monitoring site were defined by the INAG (Portuguese Institute of Water) technicians and they are corroborated by the region's topography and the land's drainage dynamics. We defined a neighbourhood around each water quality monitoring site in order to estimate monthly measurements of rainfall averages in the area to finally apply a stochastic interpolation method: Kriging. Although the study of the 20 sub-basins of water quality monitoring points has been performed, we only present the sub-basin of River Ave associated to the 8 water monitoring sites studied in this work: CANT, TAI, GOL, FER, VSA, RAV, STI, and PTR. Considering that A_i is the area of a sub-basin, the A_i, $i = 1, ..., 8$, region was regularly discretized by a set of given points: CANT 8, TAI 29, GOL 13, FER 29, VSA 15, RAV 7, STI 5, and PTR 32.

4.1 Kriging

A geostatistical approach/ordinary Kriging method was chosen with the main goal of identifying models which estimate monthly average rainfall in a sub-basin associated with a given water quality monitoring site where there are no observed values (Nicolau and Rodrigues 2000). These estimates are based on rain gauges located in their respective drainage areas. We considered that the values of precipitation recorded throughout time are approximate replicas of the same process, which is valid if the temporal correlation is weak. For this reason, in this study, the observed values of precipitation throughout the years were separated according to the twelve months of the year. For every month of the year, the spatial continuity of precipitation was analyzed to take into account the temporal component (the several months observed between 1931 and 2009).

The modelling of rainfall for each month (1931-2009) reduces significantly the well-known right skewed of this type of data. No-transformation of rainfall data was also considered in several works, for instance in Mirás-Avalos et al. (2007), in which is performed ordinary Kriging for mapping monthly rainfall data in Galicia (NW Spain), close to River Ave's hydrological basin (NW Portugal).

One objective of spatial data analysis is to predict the value $Z(s_0)$ of an observation at an unsampled site s_0 from the data $Z(s_1), ..., Z(s_n)$ at the sampled sites $s_1, ..., s_n$. Kriging is perhaps the most popular approach to spatial prediction (Cressie 1989; Rathbun 1998). In Kriging, in our case, the spatial interpolation is obtained by a linear combination of the observed values of the 19 known points s_j (meteorological stations), $j = 1, ..., 19$ and $Z_t(s_0) = \sum_{j=1}^{19} \lambda_j Z_t(s_j)$, $t = 1, 2, ..., T$ (months). If we would like to have monthly measurements, $Z_t(s)$, at any given location s in a certain area A (in particular, in our case, a sub-basin area A_i, $i = 1, ..., 20$ associated to the

20 water quality monitoring sites), we could obtain the monthly mean area precipitation through the integral

$$Z_t(A) = \frac{1}{|A|}\int_A Z_t(s)ds$$

where the integral is in the mean square sense and $|A|$ stands for the area of A. The Kriging estimator evaluates this integral as a linear combination of the measurements of the 19 available monitoring sites,

$$\hat{Z}_{t^*}(A) = \sum_{j=1}^{19} \lambda_{A,j} Z_{t^*}(s_j)$$

with the coefficients vector of $\lambda = (\lambda_{A,1}, \lambda_{A,2}, ..., \lambda_{A,19})^T$ satisfying the relation $\sum_{j=1}^{19} \lambda_{A,j} = 1$.

The method used for the calculation of the empirical semivariogram is the method of moments (Matheron 1963), modified for a random space-time process $\{Z(s,t) : s \in \mathbb{R}^2, t = 1, ..., T\}$. The process, in our case, will be a random spatial process $\{Z(s) : s \in \mathbb{R}^2\}$ which is isotropic and second-order stationary. The sampling processes are $\{Z_t(s_i) : i = 1, ..., n, t = 1, ..., T\}$, i.e. data collected at n point locations on A ($n = 19$ meteorological monitoring points) in the region considered in $\mathbb{R}^2$, for T equally spaced time instants (T months). The semivariogram is obtained by averaging the several empirical semivariograms for each time instant, a procedure that corresponds to considering the space-time process as a collection of T independent temporal replicates of a purely spatial process $\{Z(s) : s \in A\}$, in which case the purely spatial semivariogram $\hat{\gamma}_Z(h)$ characterizes all the space-time variability. Under such hypothesis, two realizations corresponding to two different but close time instants may differ substantially, since they are independent, although their spatial variability pattern remains the same (Kyriakidis and Journel 1999; Severino and Alpuim 2005). This is a usual mode of estimating the semivariogram in meteorological applications. Henceforth, the final estimator, the empirical semivariogram, is given by

$$\hat{\gamma}_Z(h \mid l) = \frac{1}{2T|N(h|l)|} \sum_{t=1}^{T} \sum_{(i,j)\in N(h|l)} [(Z_t(s_i) - Z_t(s_j)]^2 \qquad (1)$$

with $N(h|l) = \{(i,j) : \|s_i - s_j\| - \|h\| \leq l; 1 \leq i \leq j \leq n\}$ and $|N(h|l)| = \#N(h|l)$.

The models of spatial continuity (a model for each of the 12 months of the year), as were inferred from the monthly precipitation values, assume the hypothesis of homogeneity of the processes in the region under study (for a certain fixed time t^*). Under this hypothesis, two observations in the same location but in different times are independent and the spatial variability pattern remains the same. The empirical semivariograms of the 12 months were obtained by using the estimator defined in (1). For each month, the empirical semivariograms were calculated for the tolerance l in order to define the

Table 2 Values of the empirical semivariogram estimates in June: $N^*(\|h\|)$ represents the number of observation pairs from which the semivariogram estimate $\hat{\gamma}_Z(\|h\|)$ was obtained

$(\|h\|)$	$N^*(\|h\|)$	$\hat{\gamma}_Z(\|h\|)$	$(\|h\|)$	$N^*(\|h\|)$	$\hat{\gamma}_Z(\|h\|)$
0.00	770	0.000	28.50	504	660.364
4.50	116	457.589	31.50	271	624.402
7.50	397	564.012	34.50	150	628.204
10.50	409	513.998	37.50	233	795.387
13.50	327	487.421	40.50	126	887.671
16.50	578	569.138	45.50	146	1028.605
19.50	463	551.720	50.80	120	949.893
22.50	617	632.481	59.80	123	1539.302
25.50	195	582.702			

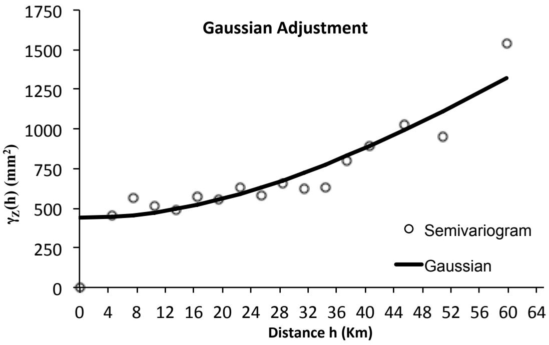

Fig. 4 Graphical representation of the estimated semivariogram with the Gaussian model adjustment

sets $N(h)$ and, therefore, the number of data pairs that are needed to estimate $\hat{\gamma}_Z(h \mid l)$. The only case presented here is the most efficient model using $l = 3$ Km tolerance. For instance, of the 12 months of the year it is only presented the semivariogram corresponding to the month of June (Table 2 shows the resulting estimates of the empirical semivariogram for this month). The least squares adjustments to several stationary models have been performed with an additional condition, enforcing the adjusted variances (the models sill value) to be equal to the empirical variance, since this is known to be the best estimator. The semivariogram model that has best performed has been the Gaussian with a nugget effect, for June in particular, as shown in Table 3. The graphical representation of the semivariogram model is shown in Figure 4. The estimated semivariogram model to describe the spatial continuity of the process in June is postulated in Eq. (2):

$$\gamma_Z(h) = \begin{cases} 0, & h = 0 \\ 440.537 + 96.5\left(1 - exp\left(-\left(\frac{\|h\|}{2770.083}\right)^2\right)\right), & h \neq 0 \end{cases} . \quad (2)$$

Table 3 Least squares adjustment results obtained with the several models: SSE denotes the residual sum of squares (June)

Model	SSE	Nugget effect	Range	Sill
Exponential	2.847×10^5	278.680	168.959	2931.940
Gaussian	1.250×10^5	440.537	96.500	2770.083
Rational Quadratic	1.529×10^5	432.931	90.644	2777.689
Spherical	2.508×10^5	290.660	286.260	2919.961

Table 4 Area precipitation estimates (in *mm*) in the neighbourhood of the water quality monitoring site of Golães during the month of June

1998	1999	2000	2001	2002	2003	2004	2005
30.86	19.40	14.26	6.66	38.47	50.65	14.34	8.21
		2006	2007	2008	2009		
		29.82	98.74	29.94	61.15		

In order to assess the quality of the semivariogram fitting, we performed a cross-validation procedure as follows: we selected a given rain gauge monitoring site at, say, s_0, based on data from the other 18 sites, then we fitted new semivariograms and estimated the ordinary Kriging (point) to obtain point estimates of $Z_t(s_0)$ across time, and finally we evaluated the corresponding residuals (differences between estimated and true values of $Z_t(s_0)$). This procedure was repeated for each of the 19 monitoring sites.

As mentioned above, the area A_i of the sub-basin associated to Golães was discretized by a set of 13 points in the region A_i included in the area A of the River Ave's hydrological basin. The estimator of the mean area precipitation $\hat{Z}_{t^*}(A_i)$ is a linear combination of the values observed in the rain gauge monitoring sites which influence the total drained area in month t^*, that is, $\hat{Z}_{t^*}(A_i) = \sum_{j=1}^{19} \lambda_{A_i,j} Z_{t^*}(s_j)$, and the value of ordinary Kriging estimation error variance is $\sigma^2_{ok} = 397.891$.

Table 4 shows the mean area precipitation estimates for the water monitoring site of Golães for the month of June, during the specific period of 1998-2009 that is relevant for this study.

4.2 Hydro-meteorological factor computation

The precipitation amount of the sub-basin A_i associated to the water monitoring site i, with $i = 1, 2, ..., 8$, in month t is estimated by $p_t^{*(i)} = \widehat{Z}_t(A_i) \times a_i$, where a_i is the sub-basin's area in Km^2. For re-scale purposes, it is considered the proportional value $p_t^{(i)} = p_t^{*(i)} \times 10^{-3}$. This construction reflects that larger areas have larger amounts of drained precipitation. On the other hand, if the goal is to obtain a prediction model for DO concentration in a month t, the hydro-meteorological factor should not incorporate the precipitation amount of the current month, but only the past information. Indeed, the precipitation effect on the river flow not only depends on the recent rainfall but also on the previous months rainfall.

Thus, it is considered a hydro-meteorological factor based on the precipitation amount of the sub-basin in time $t-1$ and $t-2$. For simplicity, for each time t and for cluster i is taken a convex linear combination of the values $p_{t-1}^{(i)}$ and $p_{t-2}^{(i)}$. Moreover, in order to attenuate extremes values, it is taken the logarithm of the convex linear combination, i.e.,

$$h_t^{(i)} = \log\left(k_1 p_{t-1}^{(i)} + (1-k_1)p_{t-2}^{(i)}\right).$$

The constant k_1 was found by numerically maximizing the linear correlation coefficient between the hydro-meteorological factor and the DO concentration variable. This procedure indicated the value $k_1 = 0.7$. Thus, the hydro-meteorological factor is taken as follows,

$$h_t^{(i)} = \log\left(0.7\, p_{t-1}^{(i)} + 0.3\, p_{t-2}^{(i)}\right).$$

5 The linear state-space model

In order to model a water quality variable Y_t based on a hydro-meteorological factor h_t and on a component as seasonality s_t (or trend), in view of the standard regression analysis this relationship can be performed by the equation $Y_t = \alpha h_t + \beta s_t + e_t$. However, this model does not accommodate changes over time neither eventual autocorrelations, even if they are weak. Indeed, water quality variables are influenced by meteorological conditions that may persist from one month to another (Alpuim and El-Shaarawi 2009). For instance, Figure 5 shows partial autocorrelations of residuals of standard linear regression of DO concentration with covariates h_t and s_t. Indeed, it is verified a significant autocorrelation in residuals.

On the one hand, the linear state-space model may be considered a standard linear regression model whose coefficients may vary over time. On the other hand, the proposed model provides the possibility of identifying and separating two structural components that are significant to explain the temporal evolution of a water quality variable. This approach is illustrated with the DO concentration in a hydrological basin with homogenous groups of water monitoring sites.

Suppose there are measures of the water quality variable at time points $t = 1, 2, ..., T$ and the river basin has k clusters of sample sites where cluster i has k_i water monitoring sites, with $i = 1, 2, .., k$. The observation equation for clusters i is:

$$\mathbf{Y}_t = [\mathbf{h}_t|\mathbf{s}_t]\,\boldsymbol{\beta}_t + \mathbf{e}_t \tag{3}$$
$$\boldsymbol{\beta}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}\left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\mu}\right) + \boldsymbol{\epsilon}_t \tag{4}$$

where the measurement equation, Eq. 3, relates to the observable water quality variable $\mathbf{Y}_t = [Y_{1,t}\ Y_{2,t} \cdots Y_{k_i,t}]'$ in the k_i sites in cluster i with the vector of unobservable variables, $\boldsymbol{\beta}_t = [\beta_{h,t}\ \beta_{s,t}]'$, called states. The $k_i \times 2$

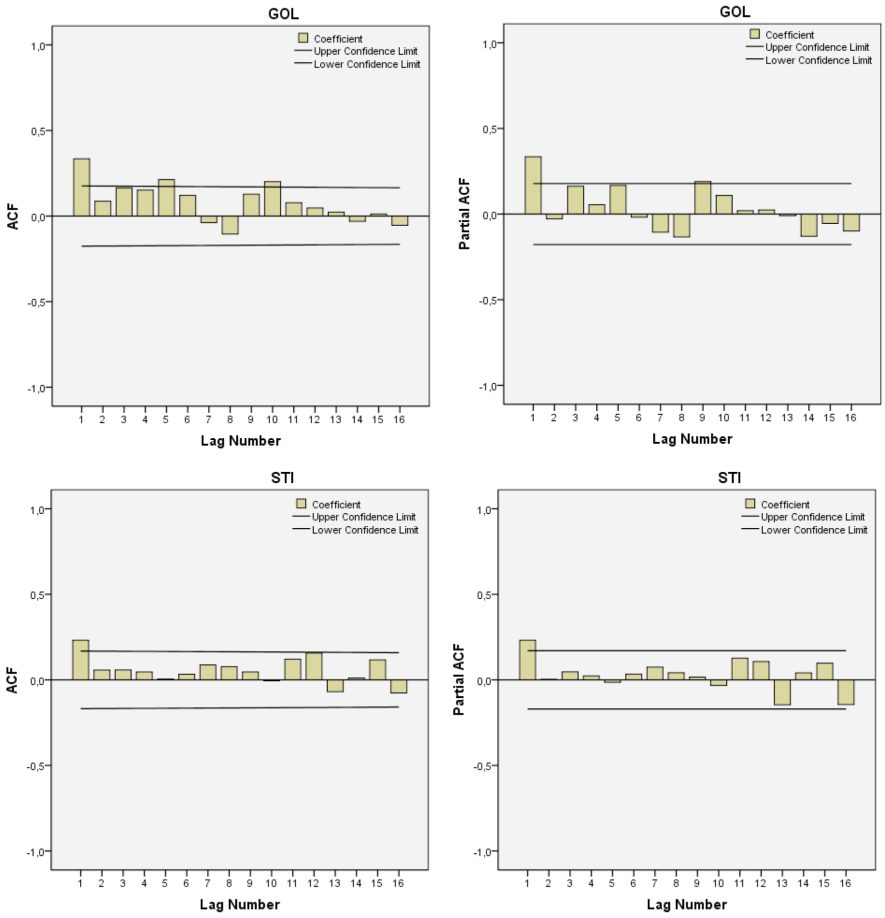

Fig. 5 PACF and ACF of residuals of the standard linear regression in *Golães* and *Santo Tirso*

matrix $\mathbf{A}_t = [\mathbf{h}_t|\mathbf{s}_t]$ is a matrix of known values and accommodates the hydro-meteorological factor and the seasonal component. Thus, matrices $\mathbf{h}_t$ and $\mathbf{s}_t$ are column matrices with the form $\mathbf{h}_t = [h_{1,t}\ h_{2,t} \cdots h_{k_i,t}]'$ and $\mathbf{s}_t = \mathbf{1}_{k_i} s_t$. For simplicity, the seasonal coefficients are taken by the mean of the monthly averages of water quality variable inside each cluster. The error term $\mathbf{e}_t$ is a white noise $k_i \times 1$ vector, called the measurement error, with a covariance matrix $\boldsymbol{\Sigma}_{\mathbf{e}}$, which may be a diagonal covariance matrix $\boldsymbol{\Sigma}_{\mathbf{e}} = diag\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_{k_i}^2\}$, for simplicity.

The state process $\{\boldsymbol{\beta}_t\}$ follows a stationary VAR(1) according to Eq. 4, the state equation, with a mean given by the 2×1 vector $\boldsymbol{\mu}$. To secure the stationarity of the state equation, it is assumed that the eigenvalues of the autoregressive matrix $\boldsymbol{\Phi}$ are inside the unit circle, i.e.,

$$|\lambda_i(\boldsymbol{\Phi})| < 1 \text{ for all } \lambda_i \text{ such that } |\boldsymbol{\Phi} - \lambda_i \mathbf{I}| = \mathbf{0}$$

and $\boldsymbol{\varepsilon}_t$ is a white noise vector with covariance matrix $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Furthermore, the noises $\mathbf{e}$ and $\boldsymbol{\epsilon}$ are serially uncorrelated, i.e., $E(\mathbf{e}_t\boldsymbol{\epsilon}_r') = \mathbf{0}$ for all t and r.

The LSS model consists of equations (3)-(4) and comprises a simplified formulation of a regression model with varying coefficients (Pagan 1980; Leybourne 2006) which includes a seasonal component. Indeed, the seasonal component could be included by adding new states in the vector state as in Bengtsson and Cavanaugh (2008). However, this approach considerably increases the complexity of the model and the matrices computations, and so the benefits are unclear. Moreover, the LSS model considered in this work enables to separate two sources of variability: one based on meteorological conditions and another on a structural component which is supposed to be more stable over time. This formulation reveals the temporal dynamic of these two factors, allowing a monthly monitoring of the water quality variable evolution. For simplicity, the seasonal coefficients s_t for each cluster are taken as known and equal to the monthly means of observations.

Usually, time series modelling takes into account, beyond seasonality, a trend component, which may be linear or otherwise. Nevertheless, the DO concentration does not present a strong trend over time, although in some water monitoring sites, mainly in more polluted clusters, there seems to be a slight linear trend. However, state-space approach is able to accommodate this behaviour because it can be interpreted as a local linear model. Indeed, the state-space model associated with the Kalman filter produces predictions of slopes in a real-time procedure, at each time, for covariates. Considering the seasonal coefficients as covariates, if there exists a trend, even weak, it is expected that the Kalman filter predictions of its slope come to be greater or lower than one. Thus, Kalman filter predictions allow signaling and monitoring possible changes in the structure evolution.

Moreover, it is known that meteorological conditions may interfere with water quality variables. Thus, the incorporation of the hydro-meteorological factor in the modelling process may contribute to separate this factor from a structural component associated with other factors, such as industries activity or an improved treatment of industrial waters. The LSS model may be an efficient real-time procedure of water quality monitoring by analyzing these two components separately.

For modelling purposes, it is necessary to predict states at each time t. As states are unobservable variables, their predictions are obtained by means of the Kalman filter algorithm (Harvey 1996). Assuming that parameters of a state-space model are known, the Kalman filter recursions give the best linear predictors to filter, forecast, and smooth the prediction of vector of states.

Let $\widehat{\boldsymbol{\beta}}_{t|t-1}$ represent the predictor of $\boldsymbol{\beta}_t$ based on the information up to time $t-1$ and let $\mathbf{P}_{t|t-1}$ be its mean square error (MSE). As the orthogonal projection is a linear estimator, the predictor for the next variable, $\mathbf{Y}_t$, is given by

$$\widehat{\mathbf{Y}}_{t|t-1} = \mathbf{A}_t\widehat{\boldsymbol{\beta}}_{t|t-1} \tag{5}$$

when, at time t, $\mathbf{Y}_t$ is available, the prediction error or innovation, $\boldsymbol{\eta}_t = \mathbf{Y}_t - \widehat{\mathbf{Y}}_{t|t-1}$, is used to update the estimate of $\mathbf{Y}_t$, through the equation

$$\widehat{\boldsymbol{\beta}}_{t|t} = \widehat{\boldsymbol{\beta}}_{t|t-1} + \mathbf{K}_t \boldsymbol{\eta}_t \tag{6}$$

where $\mathbf{K}_t$ is called the Kalman gain matrix and is given by

$$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{A}_t' \left(\mathbf{A}_t \mathbf{P}_{t|t-1}\mathbf{A}_t' + \boldsymbol{\Sigma}_{\mathbf{e}}\right)^{-1}. \tag{7}$$

Furthermore, the MSE of the update predictor $\widehat{\boldsymbol{\beta}}_{t|t}$ verifies the relationship $\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{K}_t \mathbf{A}_t \mathbf{P}_{t|t-1}$. In turn, at time t, the forecast for the state vector $\boldsymbol{\beta}_{t+1}$ is given by the equation $\widehat{\boldsymbol{\beta}}_{t+1|t} = \boldsymbol{\mu} + \boldsymbol{\Phi}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{t|t}} - \boldsymbol{\mu})$ with MSE matrix $\mathbf{P}_{t+1|t} = \boldsymbol{\Phi}\mathbf{P}_{t|t}\boldsymbol{\Phi}' + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$.

5.1 Parameters estimation

The vector of unknown parameters $\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}_{\mathbf{e}}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\}$ must be estimated from the data. In many applications, the state-space models parameters are estimated by maximum Gaussian likelihood via the Newton-Raphson method (Harvey 1996) or, more often, by the EM algorithm (Shumway and Stoffer 1982). However, environmental data may deviate from Gaussian distribution and these methods may lead to poor estimates, as it is recognized by Anagnostou and Krajewski (1998) in the context of precipitation data. Even in cases where Gaussian distribution of errors is reasonable, the use of numerical methods to maximize the log-likelihood function may be a difficult and complex task. This problem may occur because the log-likelihood function may be a complex shape with possible multiple critical points, and in these cases numerical iterative methods could not converge.

In this paper, parameters are estimated by distribution-free estimators based on the generalized method of moments. Costa and Alpuim (2010) proposed consistent distribution-free estimators for univariate state-space models which in this work are generalized to a multivariate type of model (3)-(4). Costa and Alpuim (2010) show, by using Monte Carlo studies, that the Gaussian maximum likelihood estimation produces a low rate of estimates within the space parameter in comparison to distribution-free estimators, mainly when the sample size is small. This result is a very relevant property from the practical point of view and has thus motivated its generalization. Additionally, the missing values are not a problem to the proposed estimators because they depend only on the lags between observations. More details of the distribution-free estimators construction are presented in the appendix.

Firstly, it is considered models of type (3) - (4) where the vector of states has the same dimension of the vector of the observations: for instance, 2, i.e., models with two water monitoring sites. The mean vector $\boldsymbol{\mu}$ can be easily estimated by the method of moments:

$$\widehat{\boldsymbol{\mu}} = n^{-1}\sum_{t=1}^{n} \mathbf{A}_t^{-1}\mathbf{Y}_t. \tag{8}$$

The autoregressive matrix $\boldsymbol{\Phi}$ is estimated by the covariance structure of process $\{\mathbf{A_t^{-1}Y_t}\}$ based on the autocovariance function of the process $\boldsymbol{\beta}_t$ by

$$\widehat{\boldsymbol{\Phi}} = \left(\sum_{k=1}^{\ell_{\boldsymbol{\Phi}}} \widehat{\boldsymbol{\Gamma}}_{k+1}\widehat{\boldsymbol{\Gamma}}_k^{\prime}\right)\left(\sum_{k=1}^{\ell_{\boldsymbol{\Phi}}} \widehat{\boldsymbol{\Gamma}}_k\widehat{\boldsymbol{\Gamma}}_k^{\prime}\right)^{-1} \tag{9}$$

where $\widehat{\boldsymbol{\Gamma}}_k = n^{-1}\sum_{t=1}^{n-k}\left[\left(\mathbf{A}_{t+k}^{-1}\mathbf{Y}_{t+k} - \widehat{\boldsymbol{\mu}}\right)\left(\mathbf{A}_t^{-1}\mathbf{Y}_t - \widehat{\boldsymbol{\mu}}\right)^{\prime}\right]$. The choice of $\ell_{\boldsymbol{\Phi}}$ was discussed in the original paper and is used in this paper as well. Particularly, it is suggested to take $\ell_{\boldsymbol{\Phi}} = 45, 80, 60, 50$ according to sample dimensions $n = 50, 100, 200, 500$. As the data set has samples of dimension 140, it is considered $\ell_{\boldsymbol{\Phi}} = 80$.

The state noise covariance matrix is based on relation $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \boldsymbol{\Phi\Sigma\Phi}^{\prime} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ that is valid in a VAR(1) stationary process, where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is the covariance matrix of the vector of states. To estimate $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ it is considered the estimator

$$vec\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}\right)^{\prime *} = \left(\widehat{\boldsymbol{\Gamma}}_1\widehat{\boldsymbol{\Delta}}_1^{\prime *}\right)\left(\widehat{\boldsymbol{\Delta}}_1^{*}\widehat{\boldsymbol{\Delta}}_1^{\prime *}\right)^{-1} \tag{10}$$

where matrix $\boldsymbol{\Delta}_1^*$ is obtained from the matrix $\widehat{\boldsymbol{\Delta}}_1 = \left[\mathbf{I}_4 - (\widehat{\boldsymbol{\Phi}} \otimes \widehat{\boldsymbol{\Phi}})^{\prime}\right]^{-1}\left(\mathbf{I}_2 \otimes \widehat{\boldsymbol{\Phi}}\right)^{\prime}$ summing up its two and three columns. This estimator results from (13) considering $k_{\boldsymbol{\Phi}} = 1$, following the suggestion of Costa and Alpuim (2010).

The observation noise covariance matrix $\boldsymbol{\Sigma}_{\mathbf{e}}$ is based on sample mean square error of the process $\{\mathbf{A}_t^{-1}\mathbf{Y}_t\}$, i.e., $\boldsymbol{\Gamma}_0$. Defining $\boldsymbol{\Gamma}_0$ as

$$\boldsymbol{\Gamma}_0 = \frac{1}{n}\sum_{t=1}^{n}\left[\left(\mathbf{A}_t^{-1}\mathbf{Y}_t - \boldsymbol{\mu}\right)\left(\mathbf{A}_t^{-1}\mathbf{Y}_t - \boldsymbol{\mu}\right)^{\prime}\right]$$

the estimator of $\boldsymbol{\Sigma}_{\mathbf{e}}$ is given by

$$vec\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{e}}\right)^{\prime} = n\left[vec\left(\widehat{\boldsymbol{\Gamma}}_0\right)^{\prime} - vec\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\right)^{\prime}\right]\left[\sum_{t=1}^{n}\left(\mathbf{A}_t^{-1} \otimes \mathbf{A}_t^{-1}\right)^{\prime}\right]^{-1}$$

where

$$vec\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\right) = \left[\mathbf{I}_4 - (\widehat{\boldsymbol{\Phi}} \otimes \widehat{\boldsymbol{\Phi}})\right]^{-1} vec\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}\right)$$

and with the adjustment to the covariance matrix's symmetry, similar to $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ (see the development in the appendix).

For Cluster I, the least polluted, it is possible to fit ten models, as this cluster has five water monitoring sites. Because within a cluster its elements are basically different due to the dimension of their influence area, it is considered pairs of sites with the greatest differences in their influence area, namely: *Cantelães-Ferro*, *Cantelães-Taipas*, *Taipas-Golães*, and *Vizela Santo Adrião-Taipas*. For Cluster II, the most polluted cluster, and by adopting the same methodology, two models are fitted to pairs *Riba d'Ave-Santo Tirso* and *Santo*

Table 5 Parameters estimates of linear state-space models for Clusters I and II

Cluster	$\boldsymbol{\mu}$	$\boldsymbol{\Phi}$		$\boldsymbol{\Sigma_\epsilon}$		σ_e^2				
I h_t	-0.73	0.53	0.01	1.23	-0.09	CANT	FER	TAI	GOL	VSA
s_t	1.04	1.01	0.53	-0.09	0.01	0.46	0.91	0.36	0.26	0.86
II h_t	0.02	0.27	-1.45	0.34	0.02	RAV	STI	PTR		
s_t	1.01	-0.03	0.21	0.02	0.01	1.07	1.37	0.57		

Tirso-Ponte Trofa. When for one parameter there is more than one estimate, it is considered their average.

Table 5 presents the parameters estimates for the two clusters. As expected, the expected mean value of seasonality coefficient is approximately one. The expected mean value of time varying coefficient of the hydro-meteorological factor varies around zero in Cluster II and -0.73 in Cluster I. As it be will shown further on, the DO concentration is less dependent on meteorological conditions in Cluster I than in Cluster II. Indeed, the calibration factor with a mean value of -0.73 tends to decrease the effect of rainy months that do not correspond to a significant improvement in water quality in the upstream water monitoring sites. Moreover, in more polluted sites a mean of zero in calibration factor of h_t allows incorporating a precipitation impact on the water quality, maybe for the dilution of pollutants.

Both autoregressive matrices $\boldsymbol{\mu}$ for Clusters I and II have the eigenvalues inside of unit circle that confirm the state process's stationarity. In Cluster I the covariance estimate is negative, while in Cluster II that estimate is positive. The variance estimates of the calibration factor of the hydro-meteorological covariates in Cluster I is greater than in Cluster II. This indicates that the hydro-meteorological factor has more variability in Cluster I, which can be interpreted as a less explicative covariate in comparison to Cluster II. Relatively to the individual variances σ^2, the estimation procedure shows that the higher DO concentration variability not imputed to both seasonal and hydro-meteorological components is present in sites STI and RAV, both in Cluster II.

5.2 Model's adjustment

Parameters estimates and the Kalman filter algorithm allow obtaining the predicted values for DO concentration and mainly predicted values for the calibration factors $\beta_{h,t}$ and $\beta_{s,t}$. One-step Y_t predictions (i.e. $\widehat{Y}_{t|t-1}$ in each water monitoring site) indicate the adjustment's quality and they are obtained by (5). For instance, Figure 6 shows observed values and one-step predictions of two sites in each cluster. One-step predictions fit good to the data, as it is illustrated in the same figure.

Table 6 shows the coefficients of determination (the square of the linear correlation coefficient) between observed values of DO concentration and one-step

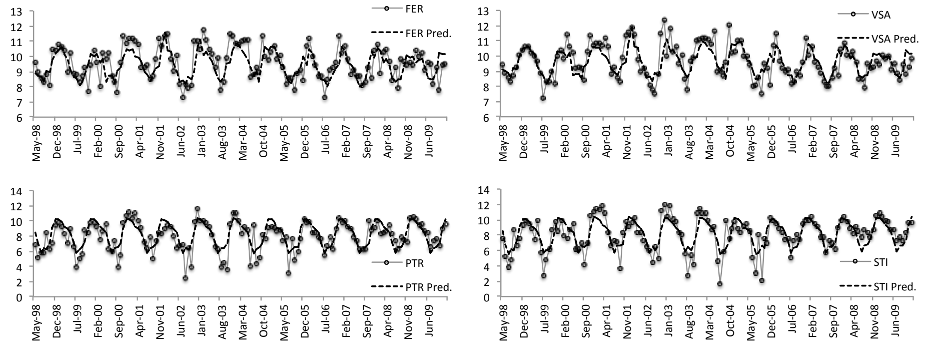

Fig. 6 Observed values of DO concentration and one-step predictions in *Ferro - FER* and *Vizela Santo Adrião - VSA* in Cluster I; *Ponte Trofa* and *Santo Tirso* in Cluster II

Table 6 Coefficients of determination of one-step predictions and filtered predictions

monitoring site	CANT	GOL	FER	VSA	TAI	RAV	STI	PTR
one-step predictions R^2	0.49	0.49	0.46	0.55	0.57	0.62	0.61	0.60
filtered predictions R^2	0.84	0.90	0.80	0.89	0.92	0.85	0.86	0.92

predictions and filtered predictions of DO concentration. It is very important to assess the adjustment of filtered predictions $\widehat{Y}_{t|t} = \mathbf{A}_t \widehat{\boldsymbol{\beta}}_{t|t}$ because one of the contributions of the proposed model is its ability to separate a structural component that accommodates a global behaviour (as the seasonality) from another component associated to hydro-meteorological conditions, represented in the hydro-meteorological covariate, which must be filtered in order to obtain the best linear predictions. In the prediction point of view, models fit good with coefficients of determination between $R^2 = 0.46$ and $R^2 = 0.62$, and the best results are achieved in water monitoring sites that are more polluted. Filtered predictions produce coefficients of determination between $R^2 = 0.80$ and $R^2 = 0.92$, which is a very good adjustment that will allow the analysis of the model components.

The main advantage of state-space models is to allow obtaining more accurate filtered predictions than the usual linear models by using the Kalman filter recursions. Indeed, linear models were adjusted to both clusters data by incorporating the seasonal and the hydro-meteorological covariates, which produce RMSE of predictions similar to the RMSE of the one-step predictions of SSM. For Cluster I, the linear model produces a RMSE of predictions of 0.84, whereas one-step predictions using the SSM produces a RMSE of 0.75. Moreover, in Cluster II the linear model predictions have a RMSE of 1.25, whereas the one-step predictions of SSM have a RMSE of 1.25. However, as mentioned before, when the filtered predictions are considered, the adjustment improves significantly since the coefficient of determination increases to values close to 1. Thus, these results show the advantages of the application of the Kalman filter equations in order to improve the predictions accuracy.

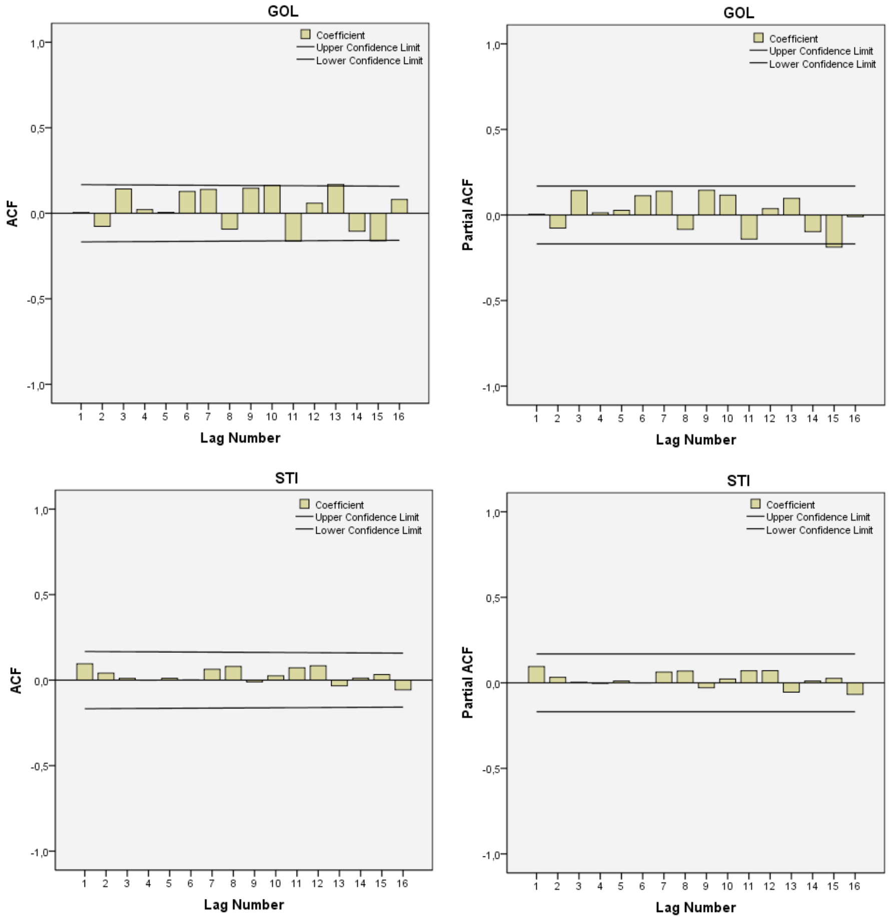

Fig. 7 Partial autocorrelations of residuals of fitted linear state-space models in *Golães* and *Santo Tirso*

In addition, the model validation should also be assessed by means of residuals analysis. Indeed, autocorrelation and partial autocorrelation functions plots (Figure 7) indicate no statistically significant serial correlation of standardized residuals, thus suggesting that any serial correlation in the data was adequately accounted for by the LSS models.

The two components included in the model (3)-(4) can be predicted by filtering the calibration factors $\beta_{h,t}$ and $\beta_{s,t}$. Indeed, the minimum mean square linear estimator of $\boldsymbol{\beta}_t$ based on observations up to and including time t is the filtered prediction $\widehat{\boldsymbol{\beta}}_{t|t}$ given by (6). Considering the parameters estimates previously obtained, the recursive equations of the Kalman filter are performed and filtered predictions of $\beta_{h,t}$ and $\beta_{s,t}$ are computed at each time t for Clusters I and II. Figure 8 represents filtered predictions for the analyzed period. As expected, the more polluted cluster (Cluster II) is more affected by hydro-

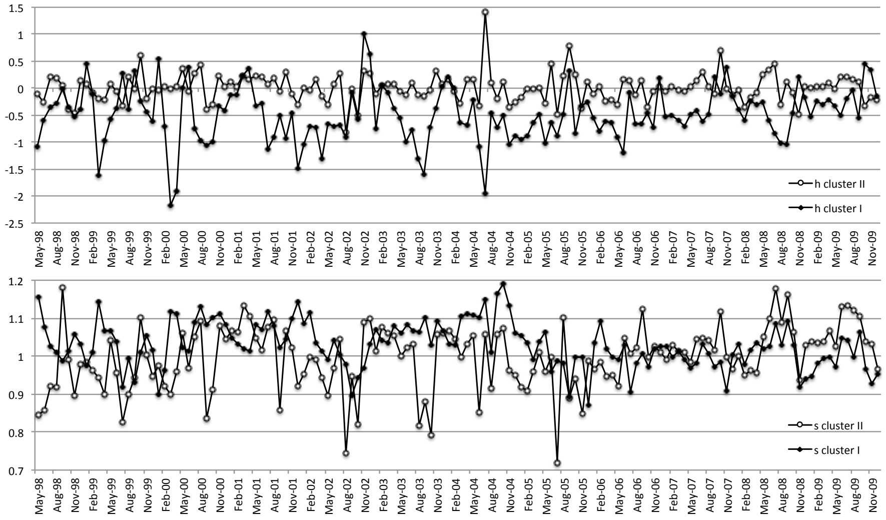

Fig. 8 Filtered predictions of the calibration factors of hydro-meteorological factor, h, and seasonality, s

meteorological conditions because its calibration factor has higher values than Cluster I. Moreover, as expected by parameters estimates in Table 5, there is a different relationship between seasonal and hydro-meteorological factors in the two clusters. Indeed, in the less polluted cluster (Cluster I) calibration factors have a negative linear correlation, while in Cluster II (the more polluted) calibration factors have a positive linear correlation. As in Cluster I, the precipitation amount has a lower impact on the DO concentration, and so it is reasonable to expect that the model tends to minimize the effect of a significant hydro-meteorological factor amount, considering that the main component of DO concentration is structural (the seasonal effect).

6 Conclusions

The analysis present in this paper allows to conclude that the hydro-meteorological factor constructed on the basis of the precipitation measure in River Ave's basin improved the prediction accuracy. Besides, the linear state-space models, associated with the Kalman filter procedure, allow to distinguish the impact of the hydro-meteorological conditions from a structural component which can incorporate exogenous factors with repercussion on the water quality variable behaviour. This modelling approach can effectively integrate these different components, and their impacts can be measured and monitored. This methodology could be further developed to better fulfil other applications requirements: for instance, other water quality variables, exogenous variables or components. Linear state-space models have a potential to outperform the usual linear regression model in terms of its ability to incorporate the temporal dynamic inherent to the water quality monitoring procedure. For instance, this approach could be used to assess water quality evolution, namely in change point detection. Indeed, the analysis of calibration factors of the structural component, such as the seasonality, could detect important changes in the water quality variability and thus attenuate the effects of the hydrometeorological conditions.

A Appendix

A.1 Distribution-free estimators for the mean and for the transition matrix

In the parameters estimation of state-space models were performed distribution-free estimators developed from the original work by Costa and Alpuim (2010). However, in that work it was proposed a distribution-free estimator for state-space models with univariate observations. Thus, a straightforward generalization of these estimators is presented in order to allow their application to a class of multivariate state-space models that largely covers the present work's needs.

To estimate the unknown parameters in the model

$$\mathbf{Y}_t = \mathbf{H}_t\boldsymbol{\beta}_t + \mathbf{e}_t \tag{11}$$

$$\boldsymbol{\beta}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}\left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\mu}\right) + \boldsymbol{\epsilon}_t \tag{12}$$

it is assumed a set of observations $\mathcal{Y}_n = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)$, and regular matrices of known constants $\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_n$ are available. The mean vector $\boldsymbol{\mu}$ can be easily estimated by the method of moments, i.e., $\widehat{\boldsymbol{\mu}} = n^{-1}\sum_{t=1}^{n}\mathbf{H}_t^{-1}\mathbf{Y}_t$.

As variables $\mathbf{Y}_t$ are not stationary, we are not under the usual conditions of the consistency of generalized method of moments. Thus, it is necessary to establish additional conditions to guarantee this consistency. By construction, the estimator $\widehat{\boldsymbol{\mu}}$ of the mean vector is unbiased, so we can guarantee its consistency by proving that $var(\widehat{\boldsymbol{\mu}}) \to \mathbf{0}$ when $n \to +\infty$, and thus establishing sufficient conditions. Covariance matrix of $\widehat{\boldsymbol{\mu}}$ is given by

$$\begin{aligned} var(\widehat{\boldsymbol{\mu}}) &= \frac{1}{n^2}\sum_{t=1}^{n}\sum_{s=1}^{n} E\left[\left(\mathbf{H_t^{-1}Y_t} - \boldsymbol{\mu}\right)\left(\mathbf{H_s^{-1}Y_s} - \boldsymbol{\mu}\right)'\right] \\ &= \frac{1}{n^2}\sum_{t=1}^{n}\sum_{s=1}^{n} E\left[\left(\boldsymbol{\beta}_t - \boldsymbol{\mu}\right)\left(\boldsymbol{\beta}_s - \boldsymbol{\mu}\right)'\right] + \frac{1}{n^2}\sum_{t=1}^{n}\mathbf{H_t^{-1}}\boldsymbol{\Sigma}_{\mathbf{e}}\mathbf{H_t'^{-1}}. \end{aligned}$$

Applying the Kronecker product $\otimes$ and the operator *vec*, we get

$$vec\left[var\left(\widehat{\boldsymbol{\mu}}\right)\right] = \frac{1}{n^2}\left[\sum_{t=1}^{n}\sum_{s=1}^{n}\boldsymbol{\Gamma}_{\boldsymbol{\beta}}(|t-s|)\right] + \frac{1}{n^2}\left[\sum_{t=1}^{n}\left(\mathbf{H_t^{-1}} \otimes \mathbf{H_t^{-1}}\right)\right] vec\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right).$$

Under the stationarity conditions of process $\{\boldsymbol{\beta}_t\}$, the first parcel is an O_p, seeing that $\sum_{k=-\infty}^{+\infty}\boldsymbol{\Gamma}(k) < \infty$, (e.g., Hamilton, 1994, p. 279). To guarantee that

$$\frac{1}{n^2}\left[\sum_{t=1}^{n}\left(\mathbf{H_t^{-1}} \otimes \mathbf{H_t^{-1}}\right)\right] vec\left(\boldsymbol{\Sigma}_{\mathbf{e}}\right) \overset{n\longrightarrow+\infty}{\longrightarrow} \mathbf{0}$$

it is sufficient to admit the additional condition $|h^{-1}_{t,(i,j)}| < c$ for all $t = 1, 2, .., i, j = 1, 2, ..., m$ and for some positive constant c, where $h^{-1}_{t,(i,j)}$ represents the (i, j) element of $\mathbf{H_t^{-1}}$ matrix.

The autoregressive matrix $\boldsymbol{\Phi}$ is estimated by means of covariance structure of process $\{\mathbf{H_t^{-1}Y_t}\}$. We see that

$$\begin{aligned}\boldsymbol{\Gamma}_{\mathbf{H^{-1}Y}}(k) &= E\left[\left(\mathbf{H^{-1}_{t+k}Y_{t+k}} - \boldsymbol{\mu}\right)\left(\mathbf{H_t^{-1}Y_t} - \boldsymbol{\mu}\right)'\right]\\ &= \boldsymbol{\Gamma}_{\boldsymbol{\beta}}(k) = \boldsymbol{\Gamma}_k.\end{aligned}$$

In a VAR(1) process, the relation $\boldsymbol{\Gamma}_k = \boldsymbol{\Phi}\boldsymbol{\Gamma}_{k-1}$ is valid, for $k = 1, 2, ...$ Thus, we proposed the autoregressive matrix estimator $\widehat{\boldsymbol{\Phi}}$ based on the least squares method of these equations by taking $k = 1, 2, ..., \ell_{\boldsymbol{\Phi}}$. Thus, we have

$$\widehat{\boldsymbol{\Phi}} = \left(\sum_{k=1}^{\ell_{\boldsymbol{\Phi}}} \widehat{\boldsymbol{\Gamma}}_{k+1}\widehat{\boldsymbol{\Gamma}}'_k\right)\left(\sum_{k=1}^{\ell_{\boldsymbol{\Phi}}} \widehat{\boldsymbol{\Gamma}}_k\widehat{\boldsymbol{\Gamma}}'_k\right)^{-1}$$

where $\widehat{\boldsymbol{\Gamma}}_k = \frac{1}{n}\sum_{t=1}^{n-k}\left[\left(\mathbf{H^{-1}_{t+k}Y_{t+k}} - \widehat{\boldsymbol{\mu}}\right)\left(\mathbf{H_t^{-1}Y_t} - \widehat{\boldsymbol{\mu}}\right)'\right]$.

By construction, the autoregressive matrix estimator is consistent, since $\widehat{\boldsymbol{\Gamma}}_k$ is a consistent estimator of $\boldsymbol{\Gamma}_k$. Whereas we have proposed a consistent estimator to $\boldsymbol{\mu}$, we consider that the mean vector $\boldsymbol{\mu}$ is known. To analyse the consistency of $\widehat{\boldsymbol{\Gamma}}_k$ we have

$$\begin{aligned}\widehat{\boldsymbol{\Gamma}}_k &= \frac{1}{n}\sum_{k=1}^{n-k}\left[(\boldsymbol{\beta}_{t+k} - \boldsymbol{\mu} + \mathbf{H^{-1}_{t+k}e_{t+k}})(\boldsymbol{\beta}_{\mathbf{t}} - \boldsymbol{\mu} + \mathbf{H_t^{-1}e_t})'\right]\\ &= \frac{1}{n}\sum_{k=1}^{n-k}\left[(\boldsymbol{\beta}_{t+k} - \boldsymbol{\mu})(\boldsymbol{\beta}_t - \boldsymbol{\mu})' + (\boldsymbol{\beta}_{t+k} - \boldsymbol{\mu})\mathbf{e'_tH'^{-1}_t} + \right.\\ &\left. + \mathbf{e'_{t+k}H'^{-1}_{t+k}}(\boldsymbol{\beta}_{\mathbf{t}} - \boldsymbol{\mu})' + \mathbf{e'_{t+k}H'^{-1}_{t+k}e'_tH'^{-1}_t}\right].\end{aligned}$$

Under the previously established condition, the last three parcels converge in probability to a null matrix. Indeed, by defining the second parcel as $\mathbf{A} = [A_{ij}]_{i,j=1,2,...,m}$ and, with some algebraic manipulation, we have

$$A_{ij} = \frac{1}{n}\sum_{t=1}^{n-k}\left((\boldsymbol{\beta}_{t,(i)} - \mu_i)\sum_{s=1}^{m} e_{t,(s)}h^{-1}_{t,(s,j)}\right)$$

and considering $\sigma_{e,(r,s)} = cov(e_{t,(r)}, e_{t,(r)})$, the variance is given by

$$var(A_{ij}) = \sigma^2_{\boldsymbol{\beta}_i}\sum_{r=1}^{m}\sum_{s=1}^{m}\sigma_{e,(r,s)}\left(\frac{1}{n^2}\sum_{t=1}^{n-k} h^{-1}_{t,(r,j)}h^{-1}_{t,(s,j)}\right)$$

If the additional condition $|h^{-1}_{t,(i,j)}| < c$ is valid, this parcel tends to $\mathbf{0}$ when $n \to +\infty$. In a similar way, we defined the third parcel by $\mathbf{B} = [B_{ij}]_{i,j=1,2,...,m}$ with elements given by

$$B_{ij} = \frac{1}{n}\sum_{t=1}^{n-k}\left((\beta_{t,(j)} - \mu_j)\sum_{s=1}^{m} e_{t+k,(s)}h^{-1}_{t+k,(i,s)}\right)$$

with variance

$$var(B_{ij}) = \sigma^2_{\beta_j}\sum_{r=1}^{m}\sum_{s=1}^{m}\sigma_{e,(r,s)}\left(\frac{1}{n^2}\sum_{t=1}^{n-k} h^{-1}_{t+k,(i,s)}h^{-1}_{t+k,(i,r)}\right)$$

Again, we guarantee that $B_{ij} = O_p$ through the same condition $|h^{-1}_{t,(i,j)}| < c$. As we shall see, this condition is a sufficient condition, as the last parcel also tends to a null matrix. Indeed, if we denote the last parcel as $\mathbf{C} = [C_{ij}]_{i,j=1,2,...,m}$, we have

$$C_{ij} = e_{t,(j)} \sum_{r=1}^{m} e_{t+k,(r)} h^{-1}_{t+k,(i,r)} h^{-1}_{t,(i,r)}$$

with variance given by

$$var(C_{ij}) = \sigma^2_{e_j} \frac{1}{n^2} \sum_{t=1}^{n-k} \sum_{r=1}^{m} \sum_{s=1}^{m} h^{-1}_{t+k,(i,r)} h^{-1}_{t,(i,r)} h^{-1}_{t+k,(i,s)} h^{-1}_{t,(i,s)} \sigma_{e,(r,s)}.$$

These results allow us to conclude that if $|h^{-1}_{t,(i,j)}| < c$, the estimator $\widehat{\boldsymbol{\Gamma}}_k$ is consistent to $\boldsymbol{\Gamma}$, when we replace the mean vector $\boldsymbol{\mu}$ by a consistent estimator.

A.2 Distribution-free estimators to noise variances

The estimation of covariance matrices of errors terms $\mathbf{e_t}$ and ε_t is an important and difficult step at the same time. At times, the recursive procedures applied to the obtained Gaussian likelihood estimates diverge or produce non-positive semidefined matrices. Sometimes, these problems occur when the initial solution is not as close to estimates as necessary. We propose an estimator to $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ based on covariance structure of a VAR(1) stationary process.
We know that the relation $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \boldsymbol{\Phi\Sigma\Phi}' + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is valid in a VAR(1) stationary process, or by applying the Kronecker product $\otimes$ and the operator *vec*

$$vec\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right) = \left[\mathbf{I}_{m^2} - (\boldsymbol{\Phi} \otimes \boldsymbol{\Phi})\right]^{-1} vec\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right).$$

By applying the *vec* operator to the equation $\boldsymbol{\Gamma}_k = \boldsymbol{\Phi\Gamma}_{k-1}$, with $k = 1, 2, ...$, we have:

$$\begin{aligned} vec\left(\boldsymbol{\Gamma}_k\right) &= vec\left(\boldsymbol{\Phi}^k \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right) \\ &= \left(\mathbf{I}_m \otimes \boldsymbol{\Phi}^k\right) vec\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right) \\ &= \left(\mathbf{I}_m \otimes \boldsymbol{\Phi}^k\right) \left[\mathbf{I}_{m^2} - (\boldsymbol{\Phi} \otimes \boldsymbol{\Phi})\right]^{-1} vec\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right) \end{aligned}$$

or

$$vec\left(\boldsymbol{\Gamma}_k\right)' = vec\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)' \left[\mathbf{I}_{m^2} - (\boldsymbol{\Phi} \otimes \boldsymbol{\Phi})'\right]^{-1} \left(\mathbf{I}_m \otimes \boldsymbol{\Phi}^k\right)'.$$

Note that the matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is symmetric, that is, $vec\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)'_{1,(j-1)m+i-1} = vec\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)'_{1,im+j}$ with $1 \le i \le m-1$ and $1 \le j \le i$. Thus, we constructed a line matrix $vec\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)'^*$, with $m + m(m-1)/2$ columns, that we got from $vec\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)'$ by removing the elements $vec\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)'_{1,im+j}$, with $1 \le i \le m-1$ and $1 \le j \le i$.
By applying the same methodology to the matrix $\boldsymbol{\Delta}_k$ defined as

$$\boldsymbol{\Delta}_k = \left[\mathbf{I}_{m^2} - (\boldsymbol{\Phi} \otimes \boldsymbol{\Phi})'\right]^{-1} \left(\mathbf{I}_m \otimes \boldsymbol{\Phi}^k\right)',$$

we summed the columns (two by two) with the index $im + j$ and $(j-1)m + i - 1$, with $1 \le i \le m-1$ and $1 \le j \le i$, thus obtaining a new matrix $\boldsymbol{\Delta}^*_k$ with $m + m(m-1)/2$ columns.
The estimator for $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is constructed via the least squares method applied to equations

$$vec\left(\boldsymbol{\Gamma}_k\right)' = vec\left(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\right)'^* \boldsymbol{\Delta}^*_k$$

with $k = 1, 2, ..., \ell_{\boldsymbol{\varepsilon}}$. Thus, we obtained the estimator

$$vec\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}\right)'^{*} = \left(\sum_{k=1}^{\ell_{\boldsymbol{\varepsilon}}} \widehat{\boldsymbol{\Gamma}}_k \widehat{\boldsymbol{\Delta}}_k'^{*}\right)\left(\sum_{k=1}^{\ell_{\boldsymbol{\varepsilon}}} \widehat{\boldsymbol{\Delta}}_k^{*} \widehat{\boldsymbol{\Delta}}_k'^{*}\right)^{-1}. \qquad (13)$$

The consistency of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}$ is guaranteed under the same conditions of the consistency of $\widehat{\boldsymbol{\Delta}}_k$. As we have seen, a sufficient condition for this is $|h^{-1}_{t,(i,j)}| < c$.

In order to estimate the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{e}}$, we defined

$$\boldsymbol{\Psi} = \frac{1}{n}\sum_{t=1}^{n}\left[\left(\mathbf{H}_t^{-1}\mathbf{Y}_t - \boldsymbol{\mu}\right)\left(\mathbf{H}_t^{-1}\mathbf{Y}_t - \boldsymbol{\mu}\right)'\right].$$

Therefore, we had the expectation

$$\begin{aligned} E(\boldsymbol{\Psi}) &= \frac{1}{n}\sum_{t=1}^{n} E\left[\left(\boldsymbol{\beta}_t - \boldsymbol{\mu} + \mathbf{H}_t^{-1}\mathbf{e}_t\right)\left(\boldsymbol{\beta}_t - \boldsymbol{\mu} + \mathbf{H}_t^{-1}\mathbf{e}_t\right)'\right] \\ &= \boldsymbol{\Sigma}_{\boldsymbol{\beta}} + \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{H}_t^{-1}\boldsymbol{\Sigma}_{\mathbf{e}}\mathbf{H}_t'^{-1}\right). \end{aligned}$$

By applying the *vec* operator, and with some algebraic manipulation, we got

$$vec\left(\widehat{\boldsymbol{\Sigma}}_{\mathbf{e}}\right)' = n\left[vec\left(\widehat{\boldsymbol{\Psi}}\right)' - vec\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\right)'\right]\left[\sum_{t=1}^{n}\left(\mathbf{H}_t^{-1}\otimes\mathbf{H}_t^{-1}\right)'\right]^{-1}.$$

As the matrix $\boldsymbol{\Sigma}_{\mathbf{e}}$ is symmetric, it is necessary to adopt the same procedure as in the estimation of $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. Thus, we estimated the $m + m(m-1)/2$ elements of the covariance matrix.

If we have a consistent estimator to $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, for example given by the proposed estimators to $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, the consistency of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{e}}$ boils down to the limit of variance of each element of $vec(\boldsymbol{\Upsilon}) = nvec(\widehat{\boldsymbol{\Psi}})'[\sum_{t=1}^{n}(\mathbf{H}_t^{-1}\otimes\mathbf{H}_t^{-1})']^{-1}$. The variance of the (i,j) element of $\boldsymbol{\Upsilon}$ is given by

$$n^2 a_{ij}^2 var\left[\frac{1}{n}\sum_{t=1}^{n}\left(\beta_{t,i} - \mu_i - \sum_{k=1}^{m} h^{-1}_{t,(i,k)}e_{t,k}\right)\left(\beta_{t,j} - \mu_j - \sum_{k=1}^{m} h^{-1}_{t,(j,k)}e_{t,k}\right)\right]$$

where $h^{-1}_{t,(i,j)}$ represents the (i,j) element of the matrix $\mathbf{H}_t^{-1}$ and a_{ij} the (i,j) element of the matrix $[\sum_{t=1}^{n}(\mathbf{H}_t^{-1}\otimes\mathbf{H}_t^{-1})']^{-1}$.

For simplicity, we adopt $\beta_{t,i} - \mu_i = \beta^*_{t,i}$. If we take in account that the states $\boldsymbol{\beta}_t$ are uncorrelated to noise $\mathbf{e}_s$ for all t and s, the previous expression can be decomposed into four parcels. The first parcel has the form

$$\begin{aligned} &a_{ij}^2 var\left(\sum_{t=1}^{n}\beta^*_{t,i}\beta^*_{t,j}\right) + a_{ij}^2 var\left(\sum_{t=1}^{n}\beta^*_{t,i}\sum_{s=1}^{m}h^{-1}_{t,(j,s)}e_{t,s}\right) + \\ &+ a_{ij}^2 var\left(\sum_{t=1}^{n}\beta^*_{t,j}\sum_{k=1}^{m}h^{-1}_{t,(i,k)}e_{t,k}\right) + a_{ij}^2 var\left[\sum_{t=1}^{n}\left(\sum_{k=1}^{m}h^{-1}_{t,(i,k)}e_{t,k}\sum_{s=1}^{m}h^{-1}_{t,(j,k)}e_{t,s}\right)\right] \end{aligned}$$

The first parcel can be decomposed into

$$a_{ij}^2 var\left(\sum_{t=1}^{n}\beta^*_{t,i}\beta^*_{t,j}\right) = a_{ij}^2\sum_{t=1}^{n}var(\beta^*_{t,i}\beta^*_{t,j}) + \sum_{t=1}^{n}\sum_{\substack{s=1\\ s\neq t}}^{n}cov(\beta^*_{t,i}\beta^*_{t,j}, \beta^*_{s,i}\beta^*_{s,j})$$

but we can write

$$var(\beta^*_{t,i}\beta^*_{t,j}) = cov(\beta^2_{t,i}, \beta^2_{t,j}) + \sigma^2_{\beta_i}\sigma^2_{\beta_j} - \gamma^2_{ij}.$$

In order for this parcel to be an O_p, it is sufficient to admit the additional regularity conditions, such as $cov(\beta_{t,i}\beta_{t,j}\beta_{s,i}\beta_{s,j})$ for all t and s, that do not depend on time.

The cross terms have a similar structure. For example, the first term can be computed by,

$$\begin{aligned} a^2_{ij} var\left(\sum_{t=1}^{n}\beta^*_{t,i}\sum_{s=1}^{m}h^{-1}_{t,(j,s)}e_{t,s}\right) &= a^2_{ij}\sum_{t=1}^{n} var\left(\sum_{k=1}^{m}h^{-1}_{t,(i,k)}\beta^*_{t,j}e_{t,k}\right) \\ &= a^2_{ij}\sum_{t=1}^{n}\sum_{k=1}^{m}h^{-2}_{t,(i,k)}\sigma^2_{\beta_j}\sigma^2_{e_k} \\ &= a^2_{ij}\sigma^2_{\beta_j}\sum_{k=1}^{m}\sigma^2_{e_k}\sum_{t=1}^{n}h^{-2}_{t,(i,k)}. \end{aligned}$$

So, if we admit that the elements of matrix $\mathbf{H}_t^{-1}$ are limited as $c_1 < |h^{-1}_{t,(i,j)}| < c_2$, where c_1 and c_2 are positive constants, it follows that this term is an O_p. In addition to these conditions on $h^{-1}_{t,(i,j)}$, if we ensure that the vector of error $\mathbf{e}_t$ is stationary of fourth-order, then we conclude that the last parcel of variance of the (i, j) element of $\boldsymbol{\Upsilon}$ is an O_p, too.

Thus, under the additional stationarity conditions of fourth-order on the vector of disturbances and the above restrictions on the elements of the matrices $\mathbf{H}_t^{-1}$, the proposed distribution-free estimator to $\boldsymbol{\Sigma}_{\mathbf{e}}$ is consistent.

Acknowledgements The authors would like to thank to Eng. Pimenta Machado from the Portuguese Regional Directory for the Northern Environment and Natural Resources, and to Eng. Cláudia Brandão from the Portuguese Institute of Water, for sharing their skills and experiences and for supplying the monitored data. A. Manuela Gonçalves acknowledges the financial support provided by the Research Centre of Mathematics of the University of Minho through the FCT Pluriannual Funding Program.

References

1. Alpuim T, El-Shaarawi A (2009) Modeling monthly temperature data in Lisbon and Prague. Environmetrics 20:835-852
2. Anagnostou EN, Krajewski WF, Seo DJ, Johnson ER (1998) Mean-field rainfall bias studies for WSR-88D. Journal of Hydrology Engineering 28:27-39
3. Ato AF, Samuel O, Oscar YD, Moi PA (2010) Mining and heavy metal pollution: assessment of aquatic environments in Tarkwa (Ghana) using multivariate statistical analysis. Journal of Environmental Statistics 1:1-13
4. Bengtsson T, Cavanaugh JE (2008) State-space discrimination and clustering of atmospheric time series data based on Kullback information measures. Environmetrics 19:103-121
5. Ciach GJ, Krajewski WF (2006) Analysis and modeling of spatial correlation structure of small-scale rainfall in Central Oklahoma. Adv Water Resour 29:1450-1463
6. Charles SP, Bates BC, Smith IN, Hughes JP (2004) Statistical downscaling of daily precipitation from observed and modelled atmospheric fields. Hydrol Process 18:1373-1394
7. Chokmani K, Ouarda TBMJ (2004) Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. Water Resour Res 40:1-13
8. Cressie NAC (1989) The many faces of spatial prediction. In Armstrong, M. (ed.), Geostatistics Vol.1, Kluwer, Dordrecht, 163-176

9. Costa M, Alpuim T (2010) Parameter estimation of state space models for univariate observations. J Stat Plan Inference 140:1889-1902
10. Costa M, Alpuim T (2011) Adjustment of state space models in view of area rainfall estimation. Environmetrics 22:530-540
11. Costa M, Gonçalves AM (2011) Clustering and forecasting of dissolved oxygen concentration on a river basin. Stoch Environ Res Risk Assess 25:151-163
12. De Marsily G (1986) Quantitative hydrogeology, 440 pp., Academic Press, London
13. Dirks KN, Hay JE, Stow CD, Harris D (1998) High-resolution studies of rainfall on Norfolk Island Part II: Interpolation of rainfall data. Journal of Hydrology, 208:187-193
14. Elhatip H, Hinis MA, Glbahar N (2008) Evaluation of the water quality at Tahtali dam watershed in Izmir-Turkey by means of statistical methodology. Stoch Environ Res Risk Assess 22:391-400
15. Gonçalves AM, Alpuim T (2011) Water Quality Monitoring using Cluster Analysis and Linear Models. Environmetrics 22:933-945
16. Goodrich DC, Faures J, Woolhiser DA, Lane LJ, Sorooshian S (1995) Measurement and analysis of small-scale convective storm rainfall variability. J Hydrol 173:283-308
17. Goovaerts P (2000) Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. J Hydrol 228:113-129
18. Greene AM, Robertson AW, Kirshner S (2008) Analysis of Indian monsoon daily rainfall on subseasonal to multidecadal time scales using a hidden Markov model. Quarterly Journal of Royal Meteorological Society 134:875-887
19. Harvey AC (1996) Forecasting Structural Time Series Models and The Kalman Filter. Cambridge University Press, Cambridge
20. Helena B, Pardo R, Vega M, Barrado E, Fernandez JM, Fernandez L (2000) Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga river, Spain) by principal component analysis. Wat Res 34:807-816
21. Isaaks EH, Srivastava RM (1989) Applied Geostatistics, 572pp., Oxford University Press, New York
22. Journel AG, Huijbregts, ChJ (1978) Mining Geostatistics, 600pp. Academic Press, London
23. Kokic P, Crimp S, Howden M (2011) Forecasting climate variables using a mixed-effect state-space model. Environmentrics 22:409-419
24. Kyriakidis PC, Journel AG (1999) Geostatistical space-time models: a review. Math Geol 31(6):651-684
25. Leybourne SJ (2006) Estimation and testing of time-varying coefficient regression models in the presence of linear restrictions. J Forecast 12(1):49-62
26. Lischeid J (2009) Non-linear visualization and analysis of large water quality data sets: a model-free basis for efficient monitoring and risk assessment. Stoch Environ Res Risk Assess 23:977-990
27. Liu CW, Lin KH, Kuo YM (2003) Application of factor analysis in the assessment of ground-water quality in a blackfoot disease area in Taiwan. Sci Total Environ 313:77-89
28. Machado A, Silva M, Valentim H (2010) A contribute for the evaluation of water bodies status in Northern Region. Revista Recursos Hídricos 31(1):57-63
29. Matheron G (1963) Principles of geostatistics. Econ Geol 58:1246-1266
30. Mc Kenna JE (2003) An enhanced cluster analysis program with bootstrap signficance testing for ecological community analysis. Environ Modell Softw 18:205-220
31. Mirás-Avalos JM, Paz-González A, Vidal-Vázquez E, Sande-Fouz P (2007) Mapping monthly rainfall data in Galicia (NW Spain) using inverse distances and geostatistical methods. Adv Geosci 10:51-57
32. Nicolau R, Rodrigues R (2000) Comparação de técnicas de interpolação espacial para mapeamento da precipitação máxima diária anual (krigagem utilizando a altitude com deriva externa). Documento interno do INAG 17:1261-1272
33. Oliveira RES, Lima MMCL, Vieira JMP (2005) An Indicator System for Surface Water Quality in River Basins. In The Fourth Inter-Celtic Colloquium on Hydrology and Management of Water Resources, Universidade do Minho, Guimarães, Portugal
34. Pagan A (1980) Some identification and estimation results for regression models with stochastically varying coefficients. J Econom 13:341-363
35. Rathbun SL (1998) Spatial modelling in irregularly shaped regions: Kriging estuaries. Environmetrics 9:109-129

36. Renwich JA, Mullan AB, Porteous A (2009) Statistical downscaling of New Zealand climate. Weather and Climate 29:24-44
37. Rossi RE, Mulla DJ, Journel AG, Franz EH (1992) Geostatistical tools for modelling and interpreting ecological spatial dependence. Ecol Monogr 62:277-314
38. Severino E, Alpuim T (2005) Spatiotemporal models in the estimation of area precipitation. Environmetrics 16:773-802
39. Simeonov V, Stratis JA, Samara C, Zachariadis G, Voutsa D, Anthemidis A, Sofoniou M, Kouimtzis TH (2003) Assessment of the surface water quality in Northern Greece. Wat Res 37:4119-4124
40. Shrestha S, Kazama F (2007) Assessment of surface water quality using multivariate techniques: A case study of the Fuji river basin, Japan. Environ Modell Softw 22:464-475
41. Shumway R, Stoffer D (1982) An approach to time series smoothing and forecasting using EM algorithm. J Time Ser Anal 3:253-264
42. Varol M, Sen B (2009) Assessment of surface water quality using multivariate statistical techniques: a case study of Behrimaz Stream, Turkey. Environ Monit Assess 159:543-553
43. Vega M, Pardo RE, Barrado & Debán (1998) Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water Res 32(12):3581-3592
44. Wurderlin DA, Diaz MP, Ame MV, Pesce SF, Hued AC, Bistoni MA (2001) Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia river basin (Cordoba-Argentina). Wat Res 35:2881-2894