2013 European Control Conference (ECC)
July 17-19, 2013, Zürich, Switzerland.

System Identification for Tide Prediction in the Venetian Lagoon

Francesca Parise and Giorgio Picci

***Abstract*— An assessment of a statistical model currently used for the prediction of high tides in the Venetian lagoon is presented. The model is analyzed from several points of view and is compared with state-space descriptions of smaller order which seem to provide slightly better results. Moreover the relevance of additional external inputs to the model, that is additional meteorological agents, is discussed.**

I. Introduction

High tide in Venice is a well-known phenomenon that has been of concern for a long time. High tide surges may result in flooding of large portions of the city and occasionally in disastrous damages. After major floodings of the city in the late sixties it has been realized that prediction of high tides should be based on suitable mathematical models. Since then a substantial effort has been devoted to the building and implementation of such models. Roughly speaking, tide prediction models can be divided in two classes, physical and statistical. There have been concentrated efforts in building physical/hydrodynamical models of water flows in and around the Lagoon which have resulted in a large literature, see e.g. [1], [5], [2], [3], [9], [4]. Modeling the effects of pressure and winds on the Adriatic basin and of water flows through the extremely complicated system of canals and marshes surrounding the city is indeed a titanic effort and, so far, it has not been so rewarding in terms of tide prediction. The general hint nowadays seems to be to proceed essentially by statistical modeling. In fact, tide predictions in Venice are currently made via a statistical model, although some of the relevant pressure data are refined via a physical model. The statistical model, called EXCO2, dates back to 1993 [12] and has been operational since then. Recently, accurate long term predictions of water levels are becoming of higher concern since the construction of the barriers (the "Mose") at the three main inlets of the Lido is coming to completion. There is a need for accurate long term predictions of water levels, especially of tide events bound to reach above the critical threshold (say 110 cm) designed to trigger the closure of the mobile gates of the Mose. For obvious economical reasons related to shipping scheduling and other activities in the lagoon, this prediction should be released way ahead in time. Scope of this study is the assessment and comparison of different methods and techniques for statistical model calibration and forecasting. In particular more recent model identification and prediction techniques, based on state-space models and Kalman filtering, are considered.

F. Parise is a PhD student with the Automatic Control Laboratory, ETH Swiss Federal Institute of Technology, Zurich: parisef@aut.ee.ethz.ch, G. Picci is with the Department of Information Engineering, University of Padova, Italy, picci@dei.unipd.it

II. Data analysis

The tide has two main components: the astronomical and the meteorological components. The first component is completely predictable, hence in this paper we shall only deal with the meteorological one. Most water level data mentioned in this paper, unless explicit mention of the contrary, will be actual water level minus the astronomical component.

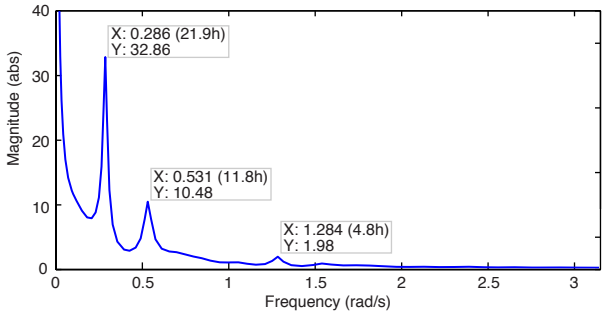

Fig. 1. Estimated spectrum of the meteorological component.

A very important physical feature of the problem is that the Adriatic sea and the gulf of Venice basin act like an oscillator cavity for sea waves. Following tidal or wind perturbations, free oscillations called seiche, (*sessa* in italian), of an approximate period of 21-22 hrs are clearly observed in the data, together with a (probable) second harmonic of about half of that period, see Fig 1. Seiche waves are very lightly damped and can, under unfavourable conditions, reach heights of 60-70 cm and add to the peak of the astronomical component with disastrous effects. The power spectral density of the (meteorological) sea level signal, estimated by fitting an AR model over a long data set, is shown in Fig 1. The seiche frequencies are clearly visible. The small peak with a period of about 5 hrs may be due to a transversal seiche along the Chioggia-Trieste direction. Estimating the bandwidth of the sea level signal will be important for prefiltering to eliminate noise. A stationarity test is run by computing moving averages of the signal over a mobile window of 720 data points (1h time-step), see Fig 2. It is evident that during autumn-winter months the meteo-induced component has a much higher impact. This is an apparent non-stationarity and a proper statistical modeling policy of this behaviour needs to be addressed; for reasons

of space this issue will have to be discussed in another paper.

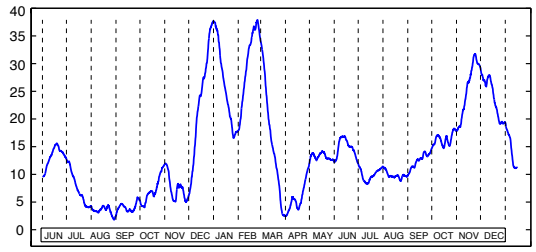

Fig. 2. Moving average of the meteo sea level data.

The effect on tide surges of meteorological agents (e.g pressure, winds, rain, etc...) is not completely understood. According to [12] " Low pressure phenomena in the Venetian basin and consequent SE winds seem to be the main exogenous causes of high tide". For this reason, the standard exogenous variables involved in the EXCO2 model are atmospheric pressures and wind strength, estimated as squared pressure gradients. The pressures are supplied by the European Centre for Medium-Range Weather Forecasts (ECMWF) in the UK. It is assumed that pressures in four key geographical locations: Alghero, Genova, Venice and Bari together with five pressure gradients across the Adriatic, namely along the joints Trieste-Ravenna, Pola-Rimini, Zara-Pescara, Spalato-Termoli, Dubrovnik-Bari, should provide an exhaustive summary of the actions of the atmospheric condition on the formation of tides. The gradients should actually be described as vectors with two components however, for simplicity, this is not done in the model currently in use. The ECMWF agency provides periodic pressure predictions which are updated daily for 180 hours ahead, with a three hours period. These predictions are essential for the purpose of the model and are actually needed also for the computation of gradient predictions. Unfortunately the wind strength estimates from pressure gradients turns out to be quite unreliable. In particular, from the comparison with real data, it seems that, although the pressure predictions are reliable, the measured winds and their effect cannot be totally explained by gradients alone. There are also questions about the accuracy of water level measurements especially regarding the bandwidth of the measuring devices. Most likely the currents at the three Lido inlets also have a role, but such data are not available. Historical hourly data used for identification are from 17/05/2009 to 31/01/2011, equal to 14,967 hourly sea level measurements at the hydrometer of "Punta della Salute". The first 10,000 data are used for model calibration (identification) and the remaining 4,967 for validation. Note that with this choice the validation is done during the autumn-winter months, i.e. when the meteorological component is more relevant.

III. ARX MODEL IDENTIFICATION

The most obvious model structure to consider is the ARX structure, the same of the existing model EXCO2. For a first analysis and comparison purposes, the same inputs of the EXCO2 model are used. The ARX model has the form

$$y(t)=c+\sum_{i=1}^{n} a_i y(t-i)+\sum_{j=1}^{9}\sum_{i=1}^{m_j} b_{j,i}u_j(t-k_j-i+1)+e(t), \tag{1}$$

where $y(t)$ is the meteorological tide while the inputs $u_1(t)$ to $u_9(t)$ are the four pressures plus five gradient inputs as listed in the previous section. The constant c is an offset due to the fact that the signals $y(t)$ and $u_j(t)$ have non zero means. For identification, the data need to be detrended by subtracting the sample means of $u(t)$ and $y(t)$, say

$$\hat{\mu}_y = \frac{1}{N_{id}}\sum_{t=1}^{N_{id}} y(t)\,, \qquad \hat{\mu}_u = \frac{1}{N_{id}}\sum_{t=1}^{N_{id}} u(t)$$

where N_{id} is the number of measurements used in the identification set. Letting

$$y'(t) = y(t) - \hat{\mu}_y, \qquad u'(t) = u(t) - \hat{\mu}_u,$$

we first estimate a detrended model

$$y'(t) = \sum_{i=1}^{n} a_i y'(t-i) + \sum_{j=1}^{9}\sum_{i=1}^{m_j} b_{j,i}u'_j(t-k_j-i+1)+e(t) \tag{2}$$

and then compute the offset by substituting the estimates into

$$c = \mu_y\left(1-\sum_{i=1}^{n} a_i\right) - \sum_{j=1}^{9}\sum_{i=1}^{m_j} b_{j,i}\mu_{u_j}, \tag{3}$$

getting

$$y(t)=\hat{c}+\sum_{i=1}^{n} \hat{a}_i y(t-i)+\sum_{j=1}^{9}\sum_{i=1}^{m_j} \hat{b}_{j,i}u_j(t-k_j-i+1)+e(t). \tag{4}$$

The structure indices $(n,\, m = [m_1 ... m_9]\,, k = [k_1 ... k_9]\,)$ need to be estimated from the data. For ARX systems this is conveniently done using the MATLAB System Identification Toolbox. The identification is done by minimizing the one-step ahead prediction error $\epsilon(t) = y(t) - \hat{y}(t|t-1)$ (PEM). Under Gaussian assumptions this estimator is known to be asymptotically equivalent to the Maximum Likelihood estimator which is consistent and has the smallest asymptotic variance [7]. As we shall see, fitting our model class to the data produces Gaussian residuals and the PEM estimate is therefore the best possible choice also for k steps ahead prediction.

A. Order selection

The first step in the identification of an ARX model is the selection of the orders. In our tide model the number of combinations $(n,\, m\,, k\,)$ with 9 different inputs is prohibitive. Hence, in a preliminary analysis, it has been assumed that m_j and k_j are the same for all the inputs and are therefore referred to simply as m and k. The implications of this assumption are discussed in Section III-B. Table I shows the best models according to the Best Fit in validation criterion,

$$FIT = 100\left(1-\frac{Var(y(t)-\hat{y}(t|t-1))}{Var(y(t))}\right),$$

where Var is the sample variance, and the well-known MDL and AIC criteria, when n is increased from 1 to 35, m from 1 to 15 and k is between 1 and 10. Higher orders are not taken into account since the corresponding model would have more than 170 parameters.

	n	m	k
Best Fit	34	12	7
MDL	33	2	7
AIC	34	2	7

TABLE I

BEST MODEL ORDERS ACCORDING TO DIFFERENT CRITERIA WHEN $n \in [1:35], m \in [1:15]$ AND $k \in [1:10]$

This table seems to suggest that the best order for the autoregressive part, n, is about the same as in the EXCO2 model[1], therefore in the following n is always fixed equal to 33. The best input delay seems to be $k = 7$. However, the pressure in Venice surely influences the level of the tide with a delay that is much smaller than 7 hours. The apparent incongruence is due to imposing the same delay $k_j \equiv k$ for all inputs. The estimated value of k should therefore be interpreted as an average value of the best delays for each input. Reasonable estimates of the average time that a mass of air takes to reach Venice from the original station range from $9-10$ hrs for Dubrovnik-Bari to zero hrs for Venice, the average of these values being between 5.4 and 6. Moreover the time needed for the pressure in Venice to influence the level of water can be assumed to be of about 1 hour, i.e. one time-step, so that the mean overall delay for the inputs is exactly $6+1=7$, as found in the previous analysis. How to choose the best value for the m parameter is also not evident from Table I since the Best Fit criterion yields $m = 12$, very similar to the value used by EXCO2, while both the AIC and MDL criteria choose $m = 2$. This difference can be partially explained by noting that the value of m has a huge impact on the overall number of parameters, $N_p = n + 9m$ while, as shown by Table II, the fit increases very slowly with m. Therefore a criterion taking into account only the best fit will choose a large value of m and tend to do overfitting.

m	2	4	6	8	10	12	14
Fit	86.42	86.39	86.42	86.45	86.45	86.46	86.43
N_p	51	69	87	105	123	141	159

TABLE II

FIT AND N_p FOR DIFFERENT VALUES OF m WHEN $n = 33$ AND $k = 7$.

B. Estimation of the delays

In this section we want to compare the performance of models when m_j and k_j are allowed to be different for different inputs. As in the previous section, the order of the autoregressive part is fixed to $n = 33$. Assume we want to test the m_j's in a range from 2 to 12 and the k_j's in a range between 1 and 10; this would lead to $(11 \cdot 10)^9 = 2.3579e{+}18$ different models, which is clearly an impossible number to simulate. Hence some assumptions on m_j and k_j are needed. In the following it has been chosen to search the values of k_j in the range intervals of Table III, which are centered around the estimates of the time that a mass of air takes to reach Venice from each station plus one hour, as detailed in Section III-A. Moreover $m_j = m$ for each input.

k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9
1	5:7	6:8	5:8	2:5	4:6	5:8	7:10	9:11

TABLE III

RANGE OF VARIABILITIES FOR k_j

Table IV shows the best values of k_j, $j : 1-9$, for several values of m, according to the Best Fit and to the MDL criterion.

m	FIT	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9
2	86.4423	1	7	6	5	3	4	5	8	11
5	86.4565	1	7	6	8	2	4	5	10	9
8	86.4631	1	7	6	8	5	5	5	8	9
12	86.4538	1	7	6	7	3	4	6	7	9
m	MDL	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9
2	3.1974	1	7	6	5	3	4	5	8	11
5	3.3268	1	7	6	8	2	4	5	10	9
8	3.4595	1	7	6	8	5	5	5	8	9
12	3.6458	1	7	6	7	3	4	6	7	9

TABLE IV

DELAYS k_j GIVING A BEST FIT (TOP) OR THE MINIMUM MDL (BOTTOM) FOR A FIXED VALUE OF m (IN THE FIRST COLUMN).

Note that, as in the previous case, the Best Fit criterion leeds to a large value of m while the MDL to a small m. Moreover it is interesting to notice that for each fixed value of m the orders selected by the two criteria are the same, since the number of parameters is the same. Finally, note that the best fit obtained with different input delays is about the same fit found with a fixed k, see Table II.

C. Analysis of the residuals

The aim of this section is to deepen the validation analysis of the best models derived in section III-A and III-B and to compare them with an ARX 33-13-1 model which mimics the structure of EXCO2. In Table V a list of the model structures discussed so far is provided. The format of the top labels is "arx n m k ", while "vd" stands for variable delays, as shown in Table IV. In addition to the previous criteria, the condition number of a companion matrix with characteristic polynomial $A(z^{-1})$ is also shown. The model with variable delays arx332vd seems to perform slightly better but all models in the table are substantially equivalent.

model	arx33131	arx34127	arx3327	arx332vd	arx338vd
FIT (id)	89.31	89.30	89.05	89.10	89.27
FIT (val)	86.40	86.48	86.44	86.38	86.46
FPE	3.1354	3.0899	3.0015	2.9998	3.0546
MDL	3.7182	3.6353	3.1986	3.1969	3.4590
AIC	1.1445	1.1297	1.0993	1.0988	1.1175
Cond	81.991	127.664	75.885	76.005	85.141

TABLE V

COMPARISON OF THE MODELS IDENTIFIED IN SECTION III-A AND III-B

[1]With the notation of formula (1) the orders of EXCO2 are $n = 33$, $m_j = 13$ and $k_j = 1$ for all j.

An important step of the validation process is to check the whiteness of the residuals, $\epsilon(t)$. This has been done, including testing the correlation of the residuals with the inputs and all the models in Table V turn out with nearly white residuals and uncorrelated inputs and residuals. The probability distribution of the residuals, estimated with a kernel method [11], turns out to be quite close to a Gaussian. This provides good evidence that a linear model should be appropriate to describe the phenomenon. A nonlinear model class (NARX) has nevertheless been attempted both on the original signals and on the residuals, but the results showed no improvement on ARX.

In conclusion, any of the ARX structures among the alternatives listed in Table V turns out to be a reasonably good model to describe the meteorological tide since the fit in validation is substantially the same, equal to 86%. Nevertheless the predictions of tide surges may occasionally be unsatisfactory. Possible causes or neglected factors are:

1) possible errors (noise) in the data
2) non-stationarity
3) influence of additional inputs (not present in the EXCO2 structure)

IV. Prefiltering

By visualizing the data it seems that the level of the meteorological tide presents high frequency components which seem to be due to noise and are of no interest for prediction. Hence we have pre-filtered the data and compared the obtained predictions to those obtainable with a model identified with the original (rough) data set. To test the effect of prefiltering we have tried an elementary non-causal average filter

$$y_s(t) = \frac{y(t-1) + y(t) + y(t+1)}{3}. \tag{5}$$

By fitting an ARX model (with orders $n = 33$, $m = 13$ and $k = 1$) to the smoothed data $y_s(t)$ instead of $y(t)$, the fit computed by averaging the one-step ahead prediction error $\hat{y}_s(t+1|t) - y_s(t+1)$ turns out to be much better than the average of 86% obtained with all the models analyzed in the previous section III-C. This can be seen comparing the fit on validation data, for three different time intervals, obtained using the standard model arx33131, identified with the measured data, and the model arx33131s described above.

	V1	V2	V3
period	(08/06/10-29/09/10)	(29/09/10-03/01/11)	$V1 \cup V2$
arx33131	75.41	88.11	86.34
arx33131s	93.14	96.31	95.96

TABLE VI
COMPARISON OF THE FIT BETWEEN MEASURED AND PREFILTERED DATA

This is a clear sign that some percentages of prediction errors are due to the presence of noise in the measured data. To implement a realistic smoothing procedure one should actually design a causal physically realizable filter, say a Butterworth or a minimum delay filter [10], with a suitable bandwidth. However since the prefiltering does not seem to be essential to alleviate the presence of macroscopic errors in the prediction of large tide surges, this refinement will be postponed to future work.

The effect of noisy data on the multi-step ahead predictions may be alleviated by a suitable assimilation procedure to optimize the initial conditions used for prediction, see e.g. [8]. Unfortunately, the results turn out to be highly dependent on the data window used for assimilation. The estimation of the initial conditions for long term prediction can be approached in a more precise and rational way by using state space models.

V. State space modeling

In an attempt to overcome the problem of noise in the data we shall resort to state-space identification. We shall describe the meteorological tide by a stationary state space model in innovation form

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + Ke(t), \\ y(t) &= Cx(t) + e(t) + d, \end{aligned} \tag{6}$$

where $x(t) \in \mathbb{R}^n$ is the state vector and n is the unknown dimension of the model, $u(t) \in \mathbb{R}^m$ is the usual input vector and $y(t) \in \mathbb{R}$ is the meteorological tide. In this case the parameters that must be identified are the matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $K \in \mathbb{R}^{n\times 1}$, $C \in \mathbb{R}^{1\times n}$ and the scalar d in a suitable canonical form.

As for the ARX model it is convenient to firstly identify a model for the zero mean vectors $y'(t)$ and $u'(t)$, and then calculate d a posteriori. Let the identified model be

$$\begin{aligned} x'(t+1) &= Ax'(t) + Bu'(t) + Ke(t), \\ y'(t) &= Cx'(t) + e(t). \end{aligned} \tag{7}$$

Note that $e(t)$ is a zero mean vector since it is the innovation process. Substituting $y'(t) = y(t) - \mu_y$, $x'(t) = x(t) - \mu_x$ and $u'(t) = u(t) - \mu_u$ yields

$$\begin{aligned} x(t+1) &= \mu_x - A\mu_x - B\mu_u + Ax(t) + Bu(t) + Ke(t), \\ y(t) &= \mu_y - C\mu_x + Cx(t) + e(t). \end{aligned} \tag{8}$$

Here μ_y and μ_u are computed as sample means of the data used for identification, while μ_x is found by solving the additional equation

$$\mu_x - A\mu_x - B\mu_u = 0. \tag{9}$$

Therefore a natural estimate of the offset d is

$$\hat{d} = \hat{\mu}_y - C\hat{\mu}_x = \hat{\mu}_y - C(I - A)^{-1}B\hat{\mu}_u. \tag{10}$$

A. Order selection

We have identifed by a subspace method (`n4sid`) several models of different state dimension and then compared their performances. Table VII shows that the fit increases very slowly when n is increased while the number of parameters increases exponentially. The MDL criterion selects low values of n, between 10 and 15. The condition number of the matrix A, denoted *Cond*, listed in the penultimate line of

Table VII, becomes very large for $n \geq 20$. The importance of the condition number of the A matrix follows from the fact that for long term predictions it will be necessary to calculate A^k for large values of the exponent k, as discussed later in Section V-B. Hence we may want to keep the condition number as small as possible. In the following analysis models of orders $n = 5$, 10 and 15 are considered.

n	5	10	15	20
FIT (id)	86.95	88.57	89.16	89.46
FIT (val)	83.65	85.49	86.10	86.37
FPE	4.1213	3.1765	2.8763	2.7310
MDL	4.3913	3.7307	3.8120	4.1420
Cond	3.1383	6.8077	5.0161	96.6417
N_p	80	210	390	620

TABLE VII

COMPARISON OF PERFORMANCE FOR DIFFERENT VALUES OF n

B. Estimation of the initial condition

Given input-output measurements up to the instant t, in order to compute the prediction from $t+1$ up to $T > t$, the one-step prediction $\hat{x}(t+1|t)$ is needed. This can be obtained by using a Kalman filter, based on the stationary identified model (6). For this model the error variances are

$$Q = \mathrm{Var}\,(Ke(t)) = \lambda^2 KK^\top, \quad R = \mathrm{Var}\,(e(t)) = \lambda^2,$$
$$S = \mathrm{Cov}\,(Ke(t), e(t)) = \lambda^2 K,$$

where $\lambda^2 = Var(e(t))$ is the innovation variance, estimated using the sample variance of the residual error

$$\lim_{N\to+\infty} \frac{\sum_{t=1}^{N} \epsilon(t)^2}{N} = \lim_{N\to+\infty} \hat{\lambda}_N^2 = \lambda^2.$$

Let $F = A - SR^{-1}C = A - KC$, $\Lambda(k) = CP(k|k-1)C^\top + R$ and $L(k) = P(k|k-1)C^\top \Lambda^{-1}(k)$ where the variance $P(k|k-1)$ satisfies the transient Riccati equation with a suitable initial condition. The equations for the prediction update step, given that the signals have non-zero means, are

$$\begin{aligned} \hat{x}(k|k) &= L(k)[C(\mu_x - \hat{x}(k|k-1)) + y(k) - \mu_y] + \\ &\quad + \hat{x}(k|k-1), \\ \hat{x}(k+1|k) &= \mu_x + F\hat{x}(k|k) + Ky(k) + Bu(k) - \\ &\quad - [F\mu_x + K\mu_y + B\mu_u]. \end{aligned} \tag{11}$$

In summary the procedure to compute $\hat{x}(t+1|t)$ is:

1) select the length l of the training interval for the Kalman filter to reach steady state;
2) set initial values $\hat{x}(t-l|t-l) = 0$ and (say) $P(t-l+1|t-l) = 100I$;
3) update the Riccati recursion and (11) until $\hat{x}(t+1|t)$.

C. Prediction

Since $E[e(t+k|t)] = 0$ for all $k \geq 1$, the state prediction vector satisfies $\hat{x}(t+k+1|t) = A\hat{x}(t+k|t) + Bu(t+k)$ for all $k \geq 1$, see also [6]. Hence, once $\hat{x}(t+1|t)$ has been estimated, the predictions can be computed as

$$\hat{x}(t+k|t) = A^{k-1}\hat{x}(t+1|t) + \sum_{i=1}^{k-1} A^{k-1-i}Bu(t+i),$$
$$\hat{y}(t+k|t) = C\hat{x}(t+k|t) + d.$$

D. Validation

In this section the performances of several state models of different order n are compared for various choices of the training interval length l. Table VIII shows the validation fit and the sample mean of the residual error, $\frac{1}{N_v}\sum_{i=1}^{N_v} |y(i+1) - \hat{y}(i+1|i)|$. As it can be seen, the best value for n is

	n \ l	30	35	40	45	50	55
Fit	5	86.05	86.09	86.12	86.15	86.18	86.19
	10	85.94	86.21	86.33	86.33	86.37	86.38
	15	85.27	85.74	86.10	86.15	86.38	86.37
Mean	5	0.027	0.031	0.031	0.027	0.030	0.031
	10	0.038	0.035	0.042	0.035	0.033	0.038
	15	0.0015	0.0023	0.0021	0.0017	0.0028	0.0039

TABLE VIII

PERFORMANCES FOR DIFFERENT VALUES OF n AND k

15 since the mean value of the residual error is one order smaller and the fit is comparable with the others. We have already seen that increasing n over 15 is not a good choice because the condition number of A would be too large. The value $l = 45$ seems to be a good compromise between a good fit and a small mean value.

E. Comparison with the ARX model

In Table IX the comparison between the mean error $\pm$ two standard deviations is shown for the ARX, the state model and EXCO2 model, as found in [12]. In particular, these values have been calculated using all the residual vector, but also using only the residual relative to a specific range of meteorological tide. From the table it seems that for surge events over 60 cm the best choice is the state-space model.

	Level	Time lead (h)			
	(cm)	1	3	6	12
State	all	0.002 ± 3.51	0.008 ± 7.95	0.014 ± 10.91	0.02 ± 11.06
	40−60	0.073 ± 4.25	1.08 ± 10.67	2.146 ± 16.91	1.27 ± 18.24
	> 60	0.863 ± 5.31	4.89 ± 12.89	8.723 ± 13.19	8.52 ± 12.17
Arx	all	0.015 ± 3.45	0.063 ± 7.76	0.123 ± 10.52	0.16 ± 10.91
	40−60	0.11 ± 3.67	0.713 ± 8.52	1.30 ± 11.55	1.34 ± 12.05
	> 60	1.10 ± 5.34	6.03 ± 13.46	11.07 ± 15.63	11.9 ± 14.21
Exco2	all	N.A.	0 ± 10	0 ± 15	0 ± 18
	40−60	N.A.	3 ± 14	8 ± 22	8 ± 25
	> 60	N.A.	6 ± 18	11 ± 22	14 ± 24

TABLE IX

MEAN ERROR $\pm$ 2 S.D. FOR MULTI STEP PREDICTIONS

VI. ADDING NEW INPUTS

Different combinations of additional inputs like wind velocity (forecasted by ECMWF), currents intensity at the inlets (roughly estimated by differentiating the astronomical tide), etc... have been investigated. In Table X there is a list of the acronyms used to indicate state models with different sets of inputs[2]. From the comparison it seems that the main advantage, visible however for long term predictions only, is for model VMVA. See Table XI.

[2]The abbreviation s.s means standard input set, i.e. pressures in Venice, Genova, Alghero and Bari + 5 gradients. The word "wind" stands for the two components of the wind velocity vector v, estimated by ECMWF, while "wind·|wind|" stands for the two components of the vector $v|v|$.

Z	no external inputs
NG	pressures in Venezia, Genova, Alghero and Bari, no gradients
S	standard input set (s.s)
V	s.s + wind in Venezia, Bari, Trieste and Dubrovnik
VMV	s.s.+ wind·|wind| in Venezia, Bari, Trieste and Dubrovnik
DA	s.s + derivative of the astronomical tide
VA	s.s + wind + derivative of the astronomical tide
VMVA	s.s.+ wind·|wind|+ derivative of the astronomical tide

TABLE X

SETS OF ADDITIONAL INPUTS

	1 step	3 steps	5 steps	10 steps
Z	84.95	64.39	51.53	42.25
NG	85.37	66.47	57.65	53.89
S	86.10	68.54	59.16	56.15
V	86.04	69.32	61.80	60.45
VMV	86.29	69.98	62.56	61.01
DA	86.43	69.98	61.36	57.47
VA	86.63	71.18	63.90	61.69
VMVA	86.58	71.14	64.10	62.10

TABLE XI

FIT OF MODELS WITH ADDITIONAL INPUTS ON VALIDATION DATA

VII. PREDICTION OF HIGH TIDE SURGES OVER 110CM

As stated before, it is important to assess the behavior of the models when the overall tide reaches the threshold of 110 cm. Fig 3 shows one such event (top).

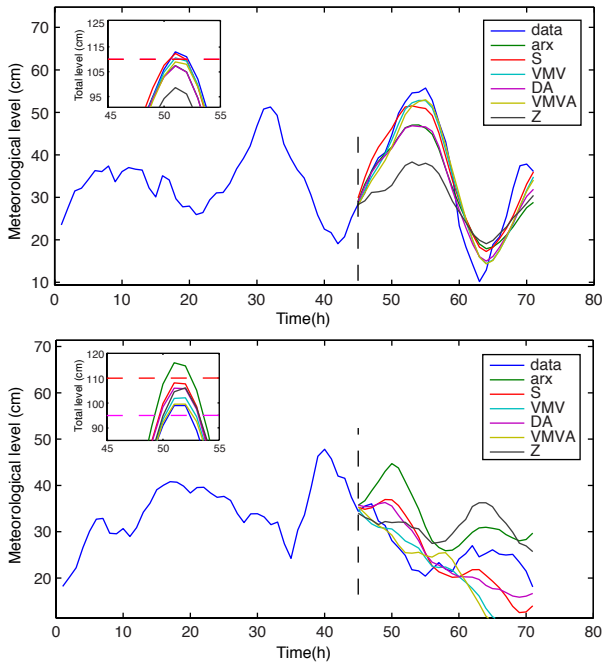

Fig. 3. Operational simulation for two events of high tide. In both figures the main plot shows the meteorological component alone, while the inset shows the total level for the same event. The first 45 data points are used to estimate the initial condition. The upper plot shows a correct prediction for an event with total level over 110 cm, the bottom one a false alarm.

Over 10 events, in 4 cases the state model with standard inputs (S) gives a closer prediction than the ARX model (A), while in 2 cases the ARX is better. In the remaining 4 cases they are approximately equally good. However it is important to note that in one case the ARX model underestimates the water level which will exceed the 110 cm threshold.

The models with wind input (VMV, VMVA) are both better than S and A in 4 cases, in 1 are worse while in the remaining 5 cases have roughly the same behavior.

A. False alarms

Another important feature for model assessment is the number of false alarms, that is how many times the model wrongly predicts values over 110 cm. Selecting all the 25 cases of total tide between 95 and 110 cm, in 9 cases we get a false alarm from at least one of the models. Fig 3 shows one of such events (bottom). In general models with wind as additional input lead to fewer false alarms: wrong predictions above 110 cm 5 times instead of 6 and wrong predictions above 115 cm 1 time instead of 3.

VIII. CONCLUSIONS

We have presented a thorough analysis of linear statistical models for the prediction of high tides in the Venetian Lagoon. It seems likely that state space models will on average behave better than the existing ARX model EXCO2, due to the use of a Kalman filter to learn the initial condition and to reduce the effect of noise in the data. Moreover it seems that the use of additional inputs, e.g. the wind velocity, could improve the prediction. Further evidence is however necessary for confirming these indications. In particular further work is needed to address the non-stationarity of the high tide phenomenon. Moreover, better forecasts of the meteorological data (e.g. the wind forecast) than those currently available would certainly improve the prediction.

REFERENCES

[1] A. Bargagli, A. Carillo, G. Pisacane, P. Ruti, M. Struglia, and N. Tartaglione, "An Integrated Forecast System over the Mediterranean Basin: Extreme Surge prediction in the northern adriatic sea," *Monthly weather review*, vol. 130, pp. 1317–1332, 2002.

[2] L. Bertotti, P. Canestrelli, L. Cavaleri, F. Pastore, and L. Zampato, "The Henetus wave forecast system in the Adriatic Sea," *Nat. Hazards Earth System Sciences*, vol. 11, pp. 2965–2979, 2011.

[3] L. Carniello, A. Defina, S. Fagherazzi, and L. D'Alpaos, "A combined wind wavetidal model for the Venice lagoon, Italy," *Journal of Geophysical Research*, vol. 110, p. F04007, 2005.

[4] L. D'Alpaos and A. Defina, "Mathematical modeling of tidal hydrodynamics in shallow lagoons: A review of open issues and applications to the venice lagoon," *Computers & Geosciences*, vol. 33, p. 476496, 2007.

[5] S. Fagherazzi, G. Fosser, L. D'Alpaos, and P. D'Odorico, "Climatic oscillations influence the flooding of Venice," *Geophysiscal reasearch letters*, vol. 32, p. L19710, 2005.

[6] G. Favier and D. Dubois, "A review of k-step-ahead Predictors," *Automatica*, vol. 26, no. 1, pp. 75–84, 1990.

[7] L. Ljung, *System identification, Theory for the user*. Prentice hall, 1999.

[8] L. Ljung and T. McKelvey, "Subspace identification from closed loop data," *Signal Processing*, vol. 52, p. 209216, 1996.

[9] T. Lovato, A. Androsov, D. Romanenkov, and A. Rubino, "The tidal and wind induced hydrodynamics of the composite system Adriatic sea/lagoon of venice," *Continental Shelf Research*, vol. 30, p. 692706, 2010.

[10] A. Mian and A. Nainer, "A fast procedure to design equiripple minimum phase fir filters," *IEEE Transactions Circuits and Systems*, vol. 29, pp. 329–331, 1982.

[11] E. Parzen, "On estimation of probability density functions and mode," *Annals of Math. Stat.*, vol. 33, pp. 1065–1073, 1962.

[12] J. Vieira, J. Fons, and G. Cecconi, "Statistical and hydrodynamic models for the operational forecasting of floods in the Venice Lagoon," *Coastal Engineering*, vol. 21, no. 4, pp. 301 – 331, 1993.