Application of Auto-Regressive (AR) analysis to improve short-term prediction of water levels in the Yangtze Estuary

Yongping Chen[1,2], Min Gan[1,2,3], Shunqi Pan[3,*], Haidong Pan[4,5], Xian Zhu[1,2], and Zhengjin Tao[1,2]

[1] State Key Laboratory of Hydrology-Water Resources & Hydraulic Engineering, Nanjing 210098, China

[2] College of Harbor, Coastal, and Offshore Engineering, Hohai University, Nanjing 210098, China

[3] Hydro-environmental Research Centre, School of Engineering, Cardiff University, Cardiff CF24 3AA, United Kingdom

[4] Key Laboratory of Physical Oceanography, Qingdao Collaborative Innovation Center of Marine Science and Technology, Ocean University of China, Qingdao, China

[5] Qingdao National Laboratory for Marine Science and Technology, Qingdao, China

Corresponding Author: Shunqi Pan (PanS2@cardiff.ac.uk)

Highlights:

1. Further identified the predictive error sources of the NS_TIDE model.
2. Established the temporal correlation of the predictive errors with AR analysis.
3. Applied AR analysis to correct the predictive errors from the NS_TIDE model.
4. Improved the short-term water level prediction of the Yangtze estuary.

Abstract

Due to the complex interaction between the fluvial and tidal dynamics, estuarine tides are less predictable than ocean tides. Although the non-stationary tidal harmonic analysis (NS_TIDE) model can account for the influence of the river discharge, the predictive accuracy of the water level in the tide-affected estuaries is yet to be improved. The results from recent studies using the NS_TIDE model in the lower reach of the Yangtze estuary showed the best root-mean-square-error (RMSE) between the predicted and measured water levels being in a range of 0.22 ~ 0.26 m. From the spectral analysis of the predictive errors, it was also found that the inaccurate description of tides in the sub-tidal frequency band was the main cause. This study is to develop a hybrid model in combination of the autoregressive (AR) analysis and the NS_TIDE model in an attempt to further improve short-term (time scale of days) water level predictions in the tide-affected estuaries. The results of the application of the hybrid model in the Yangtze estuary show a significant improvement for water level predictions in the estuary with the RMSE of 24h prediction being reduced to 0.10 ~ 0.13 m.

Keywords: water level prediction; estuarine tides; Yangtze estuary; NS_TIDE; Autoregressive model

1. Introduction

In the recent decades, estuaries have been seen the most suitable places for human settlement, agriculture, transport, and ecosystem services (Savenije, 2015). The activities of engineering development such as navigation, coastal construction, and flood protection strongly rely on accurate predictions of water levels in the estuaries, which can be vital for the safety and sustainability of the economic development in estuarine communities. In the ocean and coastal waters, water level fluctuations are mainly generated by the astronomic tides and can be predicted from the classical harmonic analysis (CHA) model such as the T_TIDE model (Pawlowicz et al., 2002) with a relatively high accuracy. However, when tides propagate in an estuary, the shallow water effect becomes significant, not only influencing their properties (amplitudes and phases), but also generating the shallow water tidal constituents (Gallo and Vinzon, 2005). In addition, the spatial variation of the estuarine geometry and the temporal change of the river discharge can further alter tidal properties in estuaries, making the characteristics of the estuarine tides more complicated. Therefore, applications of the CHA model, which is incapable of taking the influence of river discharge into account, will yield relatively inaccurate water level predictions in estuaries, particularly in the upper tidal reach. Jay (1991) developed a theory for river tide propagation in convergent channels with strong friction, and Kukulka and Jay (2003a, b) further derived the improved models describing the time-dependent tidal properties (amplitudes and phases) and tidally-averaged water levels through the nonlinear interaction of river discharge and tides. Subsequently, based on the works of Kukulka and Jay (2003a, b) and Jay et al. (2011), Matte et al. (2013) developed a non-stationary tidal harmonic analysis software package, known as NS_TIDE model. Matte et al. (2014) used the NS_TIDE model to analyse the temporal and spatial variation of tidal-fluvial dynamics in the St. Lawrence fluvial estuary. Their results showed that the model coefficients of the NS_TIDE model are location specific, but can be interpolated for other locations. Pan et al. (2018) compared the performance of the NS_TIDE model with the Empirical Model Decomposition method and their results revealed that the NS_TIDE model is less efficient in representing the sub-tidal water level fluctuations. Gan et al. (2019) explored the applicability of the NS_TIDE model to the tides in the Yangtze estuary and markedly improved its accuracy by including more sub-tidal components in the non-stationary harmonic analysis, but the predictive accuracy was found to be compromised by the additional degrees of freedom introduced in the analysis processes.

The results of Pan et al. (2018) and Gan et al. (2019) also clearly elucidated that the errors from the NS_TIDE model have strong subtidal (low-frequency) variation, with the periods longer than one or two days, indicating that the predictive errors in a short period of the past (days or months) may influence its future predictions of the low-frequency tides. In fact, the predictive errors of the NS_TIDE model, which are temporally varying, are found to have strong correlation in the time series, i.e. the future variations can be closely related to the past behaviour.

The AR analysis is expected to establish the auto-regressive (cause-and-effect) relationship between the recent and past values in the same time series, and uses the established correlation to predict the future possible values, which can be adopted to correct the tide predictions (Carbajal-Hernández et al., 2012). To account for the dynamic nature of the physical processes in engineering applications, the AR model is always linked to the Moving-Average (MA) to form the Autoregressive Moving Average (ARMA) model, which has been widely used in the areas of hydrology and oceanography. For example, Petaccia et al. (2006) used a non-linear version of the

ARMA model to forecast the sea level under high water events at Venice in Italy. Li et al. (2015) used the AR model to correct the forecast results of river discharge predicted from hydrological model and presented the potential problems exited in the application of the AR model. Similarly, Turki et al. (2015) used the ARMA model to forecast the sea level and fill the sea level gaps in oceans and coastal areas. Moreover, by considering the sea level pressure, Turki et al. (2015) improved the predictive accuracy of sea level under the surge conditions. More recently, Chen and Boccelli (2018) applied the seasonal AR model in the forecast of water demands. The ARMA models were also successfully used for other purposes such as the short-term forecast of ocean waves (Ge and Kerrigan, 2016), following the works for the spectral estimation of ocean waves (Mandal et al., 1992) and the forecast of drifting object trajectories in ocean (Minguez et al., 2012).

As shown in previous applications for a promising predictive capability, the AR model is particularly suitable for processes having a strong “memory” of the past events (Li et al., 2015). Considering the non-stationary nature of the tides in estuaries, it is proposed in this study that the AR method is to be implemented along with the NS_TIDE model as a practical tool to improve the water level predictions in an estuarine environment. Therefore, the objectives of this study are: 1) to analyse the predictive errors from the NS_TIDE model applied in the Yangtze estuary as an example to understand and establish their auto-regressive relationship; 2) to develop a practical NS_TIDE&AR hybrid model to correct the predictions from the NS_TIDE model; and 3) to examine the predictive accuracy of the NS_TIDE&AR hybrid model with the hourly measured data at several hydrometric stations along the Yangtze estuary. It should be noted that due to the regressive nature of the NS_TIDE model, the hybrid model developed in this study excludes the applicability from predicting the water levels at storm scales, where the short-term rapidly varying factors such as meteorological influence may play an important role.

2. **Model description**

2.1 Nonstationary tidal harmonic analysis (NS_TIDE) model

In the framework of CHA model for astronomical tides, the tidal amplitudes and phases are assumed to be constant. However, for the tides in estuaries, their properties (amplitudes and phases) can be strongly affected by the river discharge, and vary with both the upstream river discharge and downstream tidal range due to the nonlinear interaction between river discharge and tides. To provide the capability of achieving a better analytic accuracy than CHA model, based on the T_TIDE model (Pawlowicz et al., 2002), the NS_TIDE model takes account for both external forces (upstream river discharge and ocean tide) and their nonlinear interactions with the following equations as suggested by Matte et al. (2013, 2014):

$$\eta(t) = \eta_0 + \sum_{k=1}^{n} A_k \, cos(\sigma_k t) + B_k \sin(\sigma_k t) \tag{1}$$

where $\eta(t)$ is the water level in estuary; n is the number of tidal constituents; k is the index of tidal constituents; σ_k is the k^{th} tidal frequency; and t is time. In Eq.(1), η_0 describes the tidally-averaged water levels (frequency less than diurnal tides), commonly known as the stage model part of the NS_TIDE model and the other terms at the right-hand side represent the water level fluctuations whose frequency is equal to or higher than diurnal tides, known as the tidal-fluvial model part, as:

$$\eta_0 = c_0 + c_1 Q^{p_s}(t - t_Q) + c_2 \frac{R^{q_s}(t-t_R)}{Q^{r_s}(t-t_Q)} \tag{2}$$

$$A_k = d_{0,k}^c + d_{1,k}^c Q^{p_f}(t - t_Q) + d_{2,k}^c \frac{R^{q_f}(t-t_R)}{Q^{r_f}(t-t_Q)} \tag{3}$$

$$B_k = d_{0,k}^s + d_{1,k}^s Q^{p_f}(t - t_Q) + d_{2,k}^s \frac{R^{q_f}(t-t_R)}{Q^{r_f}(t-t_Q)} \quad (4)$$

where the subscripts *s* & *f* denote the stage model and tidal-fluvial model, respectively; (p_s, q_s, r_s) & (p_f, q_f, r_f) are unknown exponents in the stage and tidal-fluvial models, which can be determined by the iterative process using *fmincon* function in MATLAB; $Q(t)$ is the low-passed river discharge at the upstream reference location; $R(t)$ is the tidal range at the downstream reference location which takes the greater diurnal tidal range in the cases of the semi-diurnal tide regimes; t_Q & t_R are the estimated time lags of the river discharge from the upstream reference location and the tide wave from the downstream reference location propagating to a given location where water levels are modelled by the NS_TIDE model, respectively; c_i & $d_{i,k}$ (*i*= 0-2) are unknown parameters which can be determined by the iteratively reweighted least-square analysis approach (Codiga, 2011; Holland and Welsch,1977; Leffler and Jay, 2009). The time-dependent amplitude (D_k) and phase (α_k) of the k^{th} tidal constituent are calculated as,

$$\mathrm{D_k} = \sqrt{\mathrm{A_k^2} + \mathrm{B_k^2}} \text{ and } \alpha_k = tan^{-1}\left(\frac{B_k}{A_k}\right) \quad (5)$$

2.2 Auto-regressive (AR) analysis

The AR analysis is to establish the temporal correlation between the stochastic events in a time series, so to improve the predictions by taking account of the past behaviour of the variable. With the AR analysis being described in Torres et al. (2005) and Carbajal-Hernández et al. (2012), the temporal relationship of a variable, taking the difference of the water level computed by the NS_TIDE model and the measurements $\Delta\eta(t)$ as an example in this study, can be expressed as:

$$\Delta\eta(t) = \varepsilon_t + \sum_{i=1}^{p} \phi_i \Delta\eta(t-i) \quad (6)$$

where, ε_t is a random disturbance series following the stochastic process of white noise;

$\phi_1, \phi_2, \phi_3, \ldots, \phi_p$ are the autoregressive coefficients, which can be determined by the least-square method; and p is the model order of the autoregressive process.

In Eq. (6), the AR analysis order p should be sufficiently large to fairly represent the stochastic process, but a larger p will increase the degree of freedom which may on the other hand increase the instability of the model. Therefore, a criterion needs to be introduced to determine the optimal model order. To achieve the right balance of model performance and the order of freedom, in practice, the Akaike's Information Criterion (AIC) (Shibata, 1976; Torres et al., 2005) has been commonly used in determining the model order p, which can be expressed as,

$$\mathrm{AIC} = \ln[\hat{\sigma}_a^2(\mathrm{p})] + 2\frac{(\mathrm{p}+1)}{\mathrm{N}} \tag{7}$$

where $\hat{\sigma}_a^2(p)$ is the model variance and N is the number of samples. The model variance represents the model performance during the fitting period, while the model order reflects the degree of freedom of the model. With a larger p, the model variance may become smaller, but model's degree of freedom would increase. Conversely, when p is smaller, the model variance may become larger, but model's degree of freedom will reduce. Lower model's degree of freedom can make the model more stable and help deal with over-fitting problem. In determining the optimal combination of model performance and model stability, the smallest AIC number as expressed in Eq. (7) is to be sought. Smallest AIC number indicates that the AR model has the best balance between the performance and the degree of freedom of the model.

In addition, to use the AR analysis, the time series of the variable should also be ascertained to be stationary, where the time series of the variable should have no significant upward or downward variation trends and they have consistent statistical characteristics such as the mean value or variance. Therefore, the data stationarity test (Kwiatkowski et al., 1992) should be conducted prior to the model construction. In this study, the Augmented Dickey-Fuller Unit Root Test (Cavaliere and Georgiev, 2007) is used to test the stationarity of the temporal variation of the errors from the NS_TIDE model. Should the stationarity test fail, the difference method (Peters et al., 1998) will have to be applied to ensure the stationarity of the data set.

3. Study site & Field data

3.1 Yangtze estuary

Yangtze estuary is located in the middle east coast of China where the Yangtze River meets the East China Sea (**Fig. 1**). The river discharge into the Yangtze estuary has a significant seasonal variation pattern. Usually, the flood season is from May to October and the dry season starts from November and ends in April in the following year (Lu et al., 2015). Based on the recorded data, the yearly mean river discharge in dry seasons is about 10,000 ~ 20,000 m^3/s, while in flood seasons, it is about 45,000 ~ 60,000 m^3/s (Guo *et al.*, 2016). The tides at the mouth of the Yangtze estuary are predominately the semi-diurnal tides, with the mean tidal range being approximately 2.65m (Chen *et al.*, 2016).

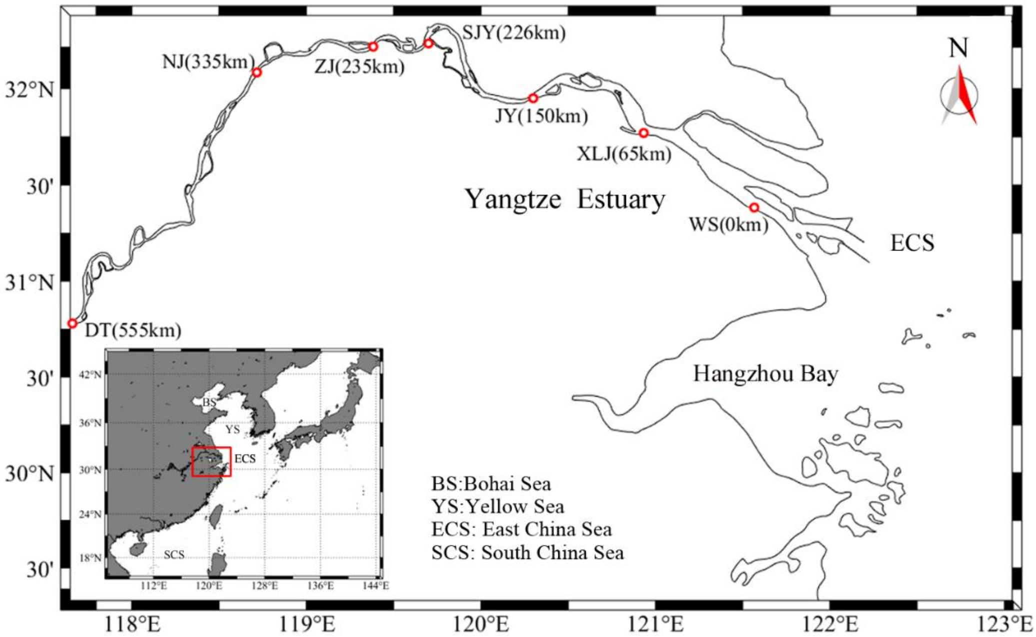

Fig. 1. Map of the low reach of Yangtze River and locations of hydrometric stations (modified from Gan et al., 2019)

3.2 Field Data

Hourly measurements of river discharge at Datong (DT) station and water levels at Nanjing (NJ), Zhenjiang (ZJ), Sanjiangying (SJY), Jiangyin (JY), Xuliujing (XLJ), Wusong (WS) stations over the period from 2014 to 2017 are available for this study. The longitudinal distance from each station to the reference location (WS station), which is the most downstream hydrologic station in the estuary, is also illustrated in **Fig. 1**. **Fig. 2** shows the time series of the measured water levels at all stations and river discharge at the most upstream station DT. It can be clearly seen that the variation of the water levels in the estuary in **Fig. 2(A)** is strongly modulated by the upstream river discharge as shown in **Fig. 2(B)**, particularly at the upper reach of the estuary, such as ZJ and NJ stations. Except for WS station, where the water levels are least affected by the river discharge, the measured water levels at all other stations in the estuary exhibit a strong non-stationarity. Therefore, WS station is used as the reference location for ocean tides together with DT station being used as the reference location for river discharge in this study.

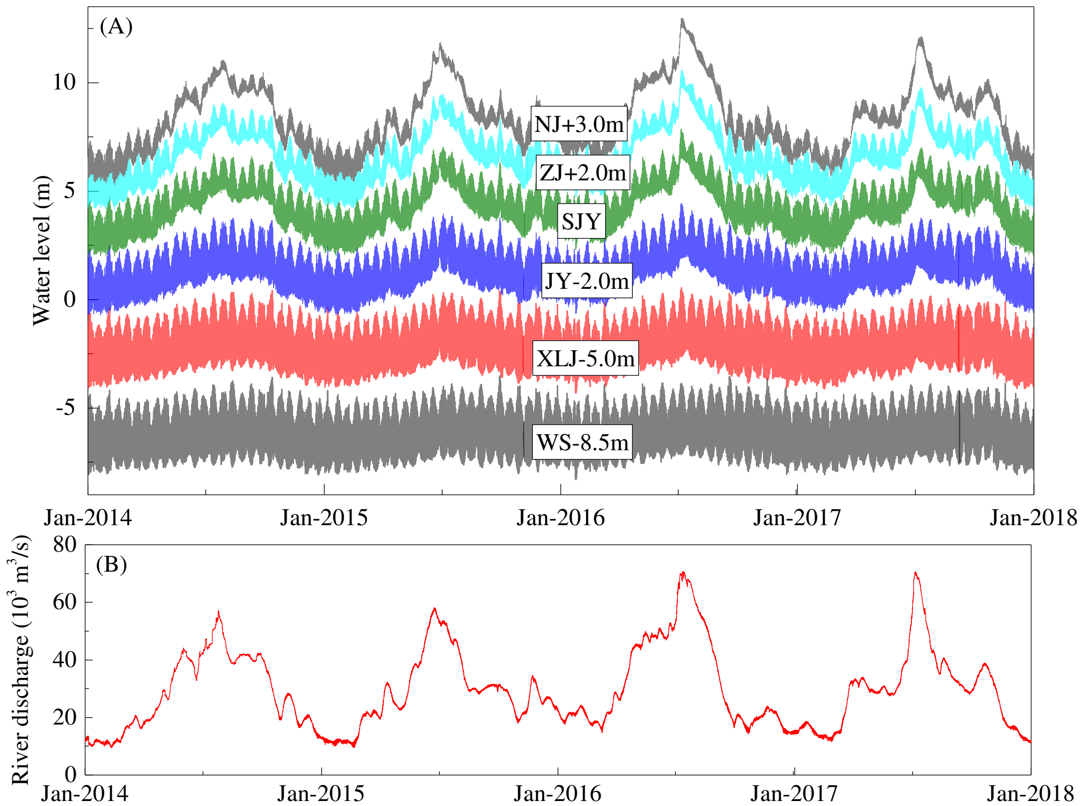

*Fig. 2. Measured water levels at 6 stations along the Yangtze estuary (A) (see **Fig. 1** for their locations); and River discharge measured at DT station (B)*

4. NS_TIDE model and Predictive errors

To create a framework for inter-comparison and assessment of the improvement of the newly proposed method, the NS_TIDE model, which is similar to that used in the work of Gan et al. (2019), is applied to the study site over the entire 4-year period of the available measurements (January 2014 - December 2017). Within the measurement period, the measured water levels over 8785 hours (i.e. 8785 hourly measurements) are used to regress the model coefficients of the NS_TIDE model. To account for the seasonality of the river discharge and dynamic nature of the tides, those coefficients are renewed regularly after a certain period of time (D hours), which is 360 hours in this study, in considering the neap-spring tidal cycles.

It should be noted that when the regression procedure can be carried out over a sufficiently long period of the measurements that covers the all (low and high) river flow conditions, updating the NS_TIDE model parameters may not be necessary as suggested by Matte et al. (2013). However, as can be seen from **Fig. 2(B),** the measurements available to this study only cover a period of low flow years (2014 - 2015) and a period of high flow years (2016 – 2017), and the regression procedure can only cover part of the entire measurement period, it therefore becomes necessary to update the model parameters regularly as aforementioned in this study to better capture the seasonal variation of the river discharge and improve the model performance. However, updating the model parameters of the NS_TIDE model sometimes may also incur discontinuity in the model parameters, but this was found to be rather minor in the present study.

To further illustrate the applicability of the NS_TIDE model in the Yangtze estuary with the proposed settings, the results from T_TIDE, NS_TIDE with theoretical exponents, and the NS_TIDE model with the optimised exponents in this study are compared. As the NS_TIDE model is developed on the frameworks of Kukulka and Jay (2003a, b), it is assumed that similar or even smaller magnitude of the tidal discharge relative to the river discharge, and a moderate estuarine shape convergence. Although at the downstream reach of Yangtze estuary, the mean tidal prism can be around10 times larger than the mean river discharge, which may partly limit the applicability of the NS_TIDE model, the iterative method to determine the exponents in Eqs. (2~4) would be the effective way to relax this requirement in the assumptions. In terms of the geometry of the estuary, the very low reach of the Yangtze estuary (from estuary mouth to JY) covers about 200 km and the channel width decreases from nearly 20 km to 3 km toward the upstream, while in the upstream reach of the estuary (JY - DT), it keeps nearly uniform width though with the presence of meanders (Guo et al. 2015). Therefore, it can be reasonably assumed that the variation of the channel width meets the requirement of the NS_TIDE model, which was also indicated in the studies of Zhang et al. (2012) and Cai et al. (2014). The results of Gan et al (2019) also illustrated the applicability of the NS_TIDE model in the Yangtze estuary with confidence. For the NS_TIDE model with theoretical exponents, the following values as suggested by Kukulka and Jay (2003a, b) are used: $(p_s = 2/3,\ q_s = 2,\ r_s = 4/3)$ & $(p_f = 1,\ q_f = 2,\ r_f = 1/2)$. From the tests carried out during the period between 2014 and 2017, the RMSE values for all 3 models are compared in **Fig. 3.** The results show that the NS_TIDE model with the iteratively optimized exponents as used in this study preforms better than the NS_TIDE model with the theoretical exponents and the classical harmonic analysis (T_TIDE) model (Pawlowicz et al. 2002), with the RMSE values being in the range of 0.20 – 0.25 m. At XLJ station, which is closest to the estuary mouth where the effect of the river discharge on the water level is expected to be the least, the classical harmonic analysis (T_TIDE) model would perform better. The results further illustrate the applicability of the NS_TIDE model in the Yangtze estuary.

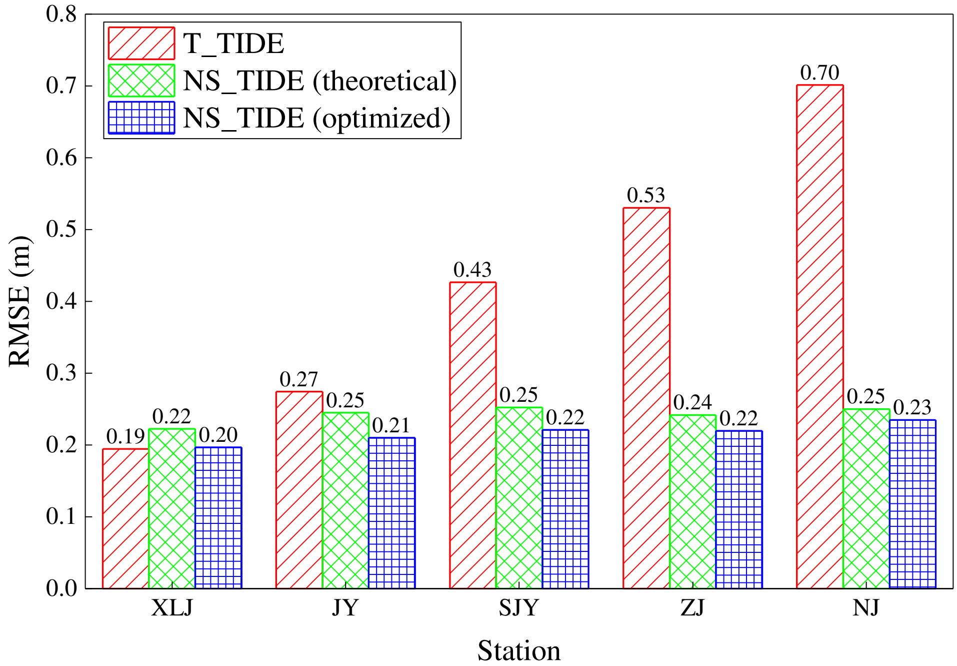

Fig. 3. Comparison of the RMSE values of the T_TIDE model and the NS_TIDE model with theoretical and iteratively optimized exponents.

With the applications of the NS_TIDE model, **Fig. 4** shows the analysed tidal amplitudes and phases at ZJ station from the NS_TIDE model. As shown in **Fig. 4**, during the period from 2014 to 2017, there are 4 flood-dry seasonal variations as indicated by the river discharge measured at DT station. The tide analysis shows a clear modulation of the seasonal river discharge variations on the tidal amplitudes and phases of M_2 and S_2 tidal constituents. The tidal amplitudes of M_2 and S_2 tidal constituents decrease with the increase of river discharge, while their phases increase with the increase of river discharge. The variation of the tidal amplitudes and phases of M_2 and S_2 tidal constituents reflects the effect of frictional dissipation and retardation of river discharge on the propagation of tidal waves. In addition, it should be noticed that the variation of the tidal amplitudes and phases of M_2 and S_2 tidal constituents presents fortnightly variation patterns. The annual and fortnightly variation cycles of the amplitudes and phases of M_2 and S_2 tidal constituents correspond to the annually-varied upstream river discharge at the DT station and fortnightly-varied (neap-spring cycle) tidal ranges at WS station, which is reflected in the Eq. (1). Comparing the decreased tidal amplitudes for M_2 and S_2 during the annual peak river discharges (**Fig. 4)** with the increased total water levels at those peaks shown in **Fig. 2**, clearly indicates that the lower frequency tidal constituents can make a considerable contribution to the total water level, which is a key aspect to be investigated in the following sections.

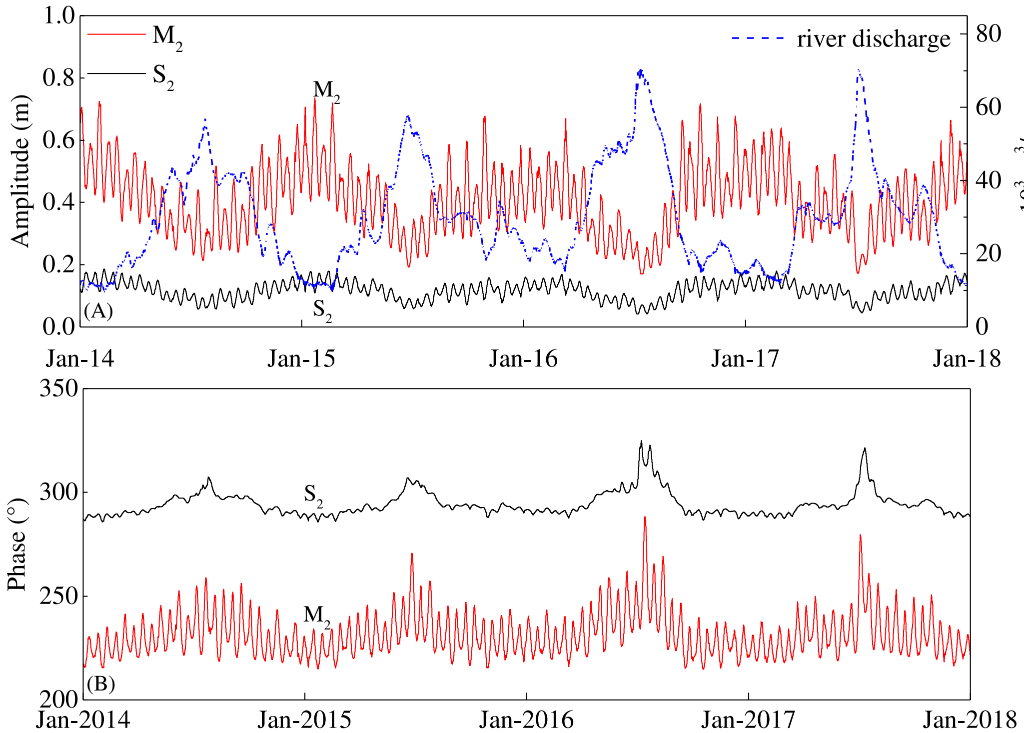

Fig. 4. The influence of the river discharge on: (A) tidal amplitudes; and (B) phases of M_2 and S_2 tidal constituents at ZJ station.

To obtain the predictive errors of the NS_TIDE model relative to the field measurements, the water levels between January 2015 and December 2017 are predicted at 5 stations: XLJ, JY, SJY, ZJ and NJ and the root-mean-square-error (RMSE) values of the predicted values relative to the measurements are calculated. **Fig. 5** shows the predictive errors of the NS_TIDE model at ZJ station superimposed on their low-passed values. The low-passed filtering process is to filter the fluctuation whose frequency is larger than 1 cycle per day (cpd). The results show that the predictive errors are mainly between ±0.5 m with the RMSE value being about 0.25 m. However, there are considerable low-frequency fluctuations of the predictive errors (**Fig. 5**), which indicates the inaccuracy of the NS_TIDE model in predicting the water levels with longer period than diurnal period. In other words, the predictive errors of the NS_TIDE model have strong variation at the subtidal frequency bands.

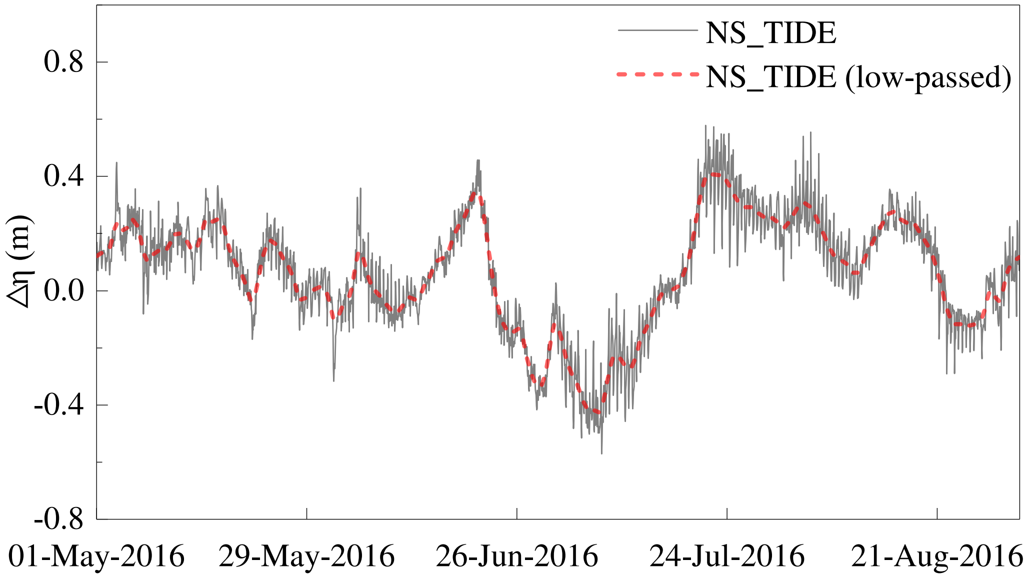

Fig. 5. Predictive errors of the NS_TIDE model and its related low-passed values at ZJ station.

To further show the energy distribution of the predictive errors of the NS_TIDE model, spectral analysis is also applied. **Fig. 6** shows the spectral energy distribution of the predictive errors of the NS_TIDE model at the 5 stations along the Yangtze estuary. In the frequency domain, the results show that the predictive errors of the NS_TIDE model has the peak spectral energy in the subtidal (D_0) band. In D_0 band, at the frequency around 0.07 cpd (neap-spring cycle), there is a general increasing trend with the decrease of frequency. This means the predictive errors of the NS_TIDE model in D_0 band partly come from inaccurately presenting the neap-spring cycles of estuarine tides. This also justifies the need of updating the model coefficients of the NS_TIDE model after each neap-spring cycle (D=360 points) adopted in this study. The spectral peaks are also found at diurnal (D_1), semi-diurnal (D_2), and quarter-diurnal (D_4) tides, but relatively smaller than the tides in D_0 band. This implies that the NS_TIDE model preforms relatively worse in modelling the tides from subtidal band than the tides in diurnal or higher frequency tidal bands.

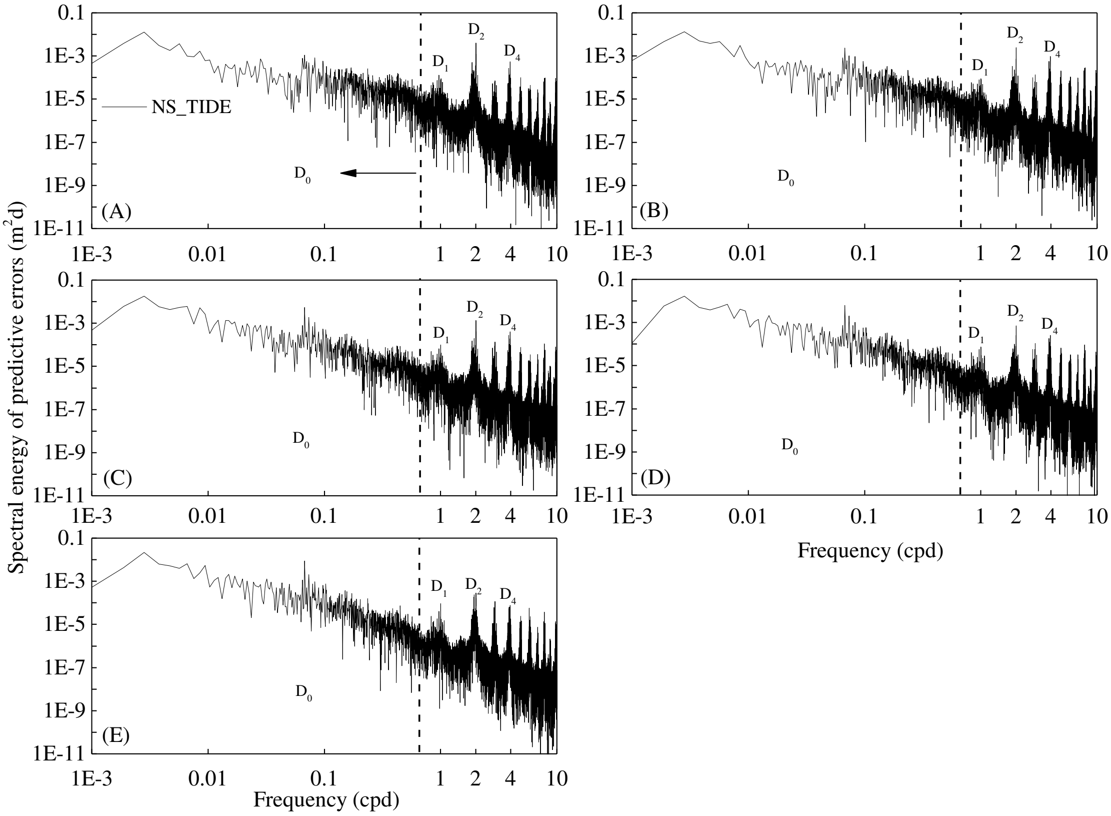

Fig. 6. Spectral energy distribution of the predictive errors of the NS_TIDE model at XLJ (A), JY (B), SJY (C), ZJ (D), and NJ (E) stations.

5. AR analysis

As shown in **Figs**. **5** and **6**, the predictive errors of the NS_TIDE model is found mainly in the subtidal (low-frequency) band. This feature therefore makes the AR analysis more suited to be applied. To this end, this study is to introduce the AR analysis through relating the temporal correlation of the predictive errors of the NS_TIDE model to the current or further predictive errors for the NS_TIDE model by treating the predictive errors as a stochastic process, so that the tide predictions from the NS_TIDE model can be improved, particularly in the low-frequency subtidal bands.

The AR analysis as illustrated in **Fig. 7** consists of two stages: Stage 1 is to determine the parameters required for the AR analysis and optimal model order from the results of the NS_TIDE model and the measurements, and Stage 2 is to apply AR analysis with the NS_TIDE model (as the NS_TIDE&AR hybrid model) to examine the improvement of the water level predictions. Specifically, in Stage 1, a period of N points (hours) of the results is used to determine the model order *p* which is optimised by the AIC criterion; and in Stage 2 the constructed AR analysis is used to correct the predictions of the NS_TIDE model over the second part of the data (say L points), which can be regarded as the test period.

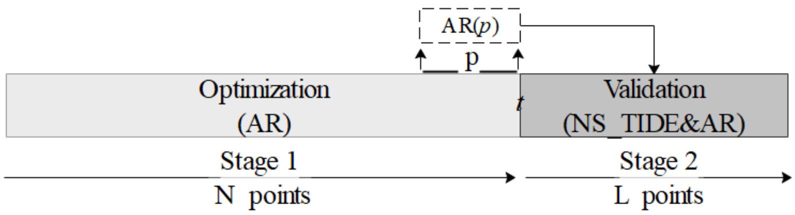

Fig. 7. Conceptual diagram of the NS_TIDE&AR hybrid model

For the AR analysis in this study, there are 2 key parameters that should be determined. One is the upper limit of the model order *p*, and the other is the number of samples *N* for the regression of the model coefficients. Usually, the model order *p* is determined from the partial autocorrelation function of time series with constant sample number. However, in this study, the sample is considered dynamically changing. Therefore, an upper limit of the model order *p* in this study is initially specified before the determination of the optimal model order *p*. The upper limit of the model order *p* is determined by preliminary numeric experiments. The determination the upper limit of p uses 720 hours data points from the NS_TIDE model, which covers a period of one month for 2 spring-neap tide cycles. Model order p is initially set to 5 and increased to 100. The optimal upper limit of p is then determined when the AIC value reaches a stable value. **Fig. 8** shows the variation of the AIC values of AR analysis with different model order p at XLJ and ZJ stations as examples. It can be seen that there are 2 sharp decreases of the AIC values at p equalling to 12 and 25. Those two locations appear to be corresponding to 2 spectral peaks around D_2 and D_1 in **Fig. 6**. Compared with **Fig. 6**, **Fig. 8** further indicates the 2 spectral peaks may be related to both semi-diurnal and diurnal tides whose periods are around 12h and 25h such as M_2 (or N_2), and O_1 tidal constituents. When p is further increased, the AIC value continues to decrease and reach to a stable state at p=40. However, when p is larger than 40, the performance of AR analysis is no longer significantly improved at both locations. Therefore, in this study, the upper limit of p at each station is set to 40.

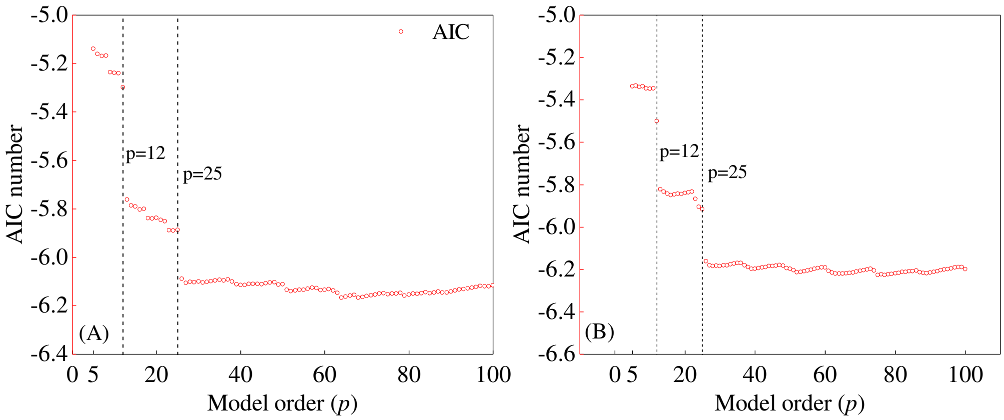

Fig. 8. Numerical experiment for determining the optimal upper limit of the model order p at: XLJ (A); and ZJ (B)

Once the upper limit of the model order p is determined, the optimum number of samples (N) is another key parameter in the AR analysis to calculate the autoregressive coefficients. For the steady cases, *N* can be regarded as constant following the works of Torres (2005) and Mirzavand and Ghazavi (2015). However, for gradually unsteady processes, such as the water levels in this study, the model's number of samples should be considered as dynamic and requires to be renewed periodically. In this study, tests are carried out with the hybrid NS_TIDE&AR model with varying number of samples (N=360-1800) over different predictive durations (L=12-48 hours). **Fig. 9(A-E)** shows the RMSE values of the predictions of the water levels from the hybrid NS_TIDE&AR model against the measurements for those tests. Overall, the RMSE values show a decreasing trend with the increase of *N* and the results become stable in almost all cases when *N* is greater than 1200 (hours). However, with a large *N*, the computational costs will also become higher and fluctuations of the RMSE values may occur at some stations due to the over-fitting. To balance the model accuracy and computational costs, the number of samples of the NS_TIDE&AR model at all stations are set to 1440 (corresponding to 1440 hours), which is equivalent to a 2-month period. **Fig. 9(F)** shows the comparison of RMSE values when N=1440 for different predictive durations, illustrating a high consistency and a slight increasing trend when the predictive duration increases from 12 to 48 hours, but all within a range of 0.08 to 0.16 m.

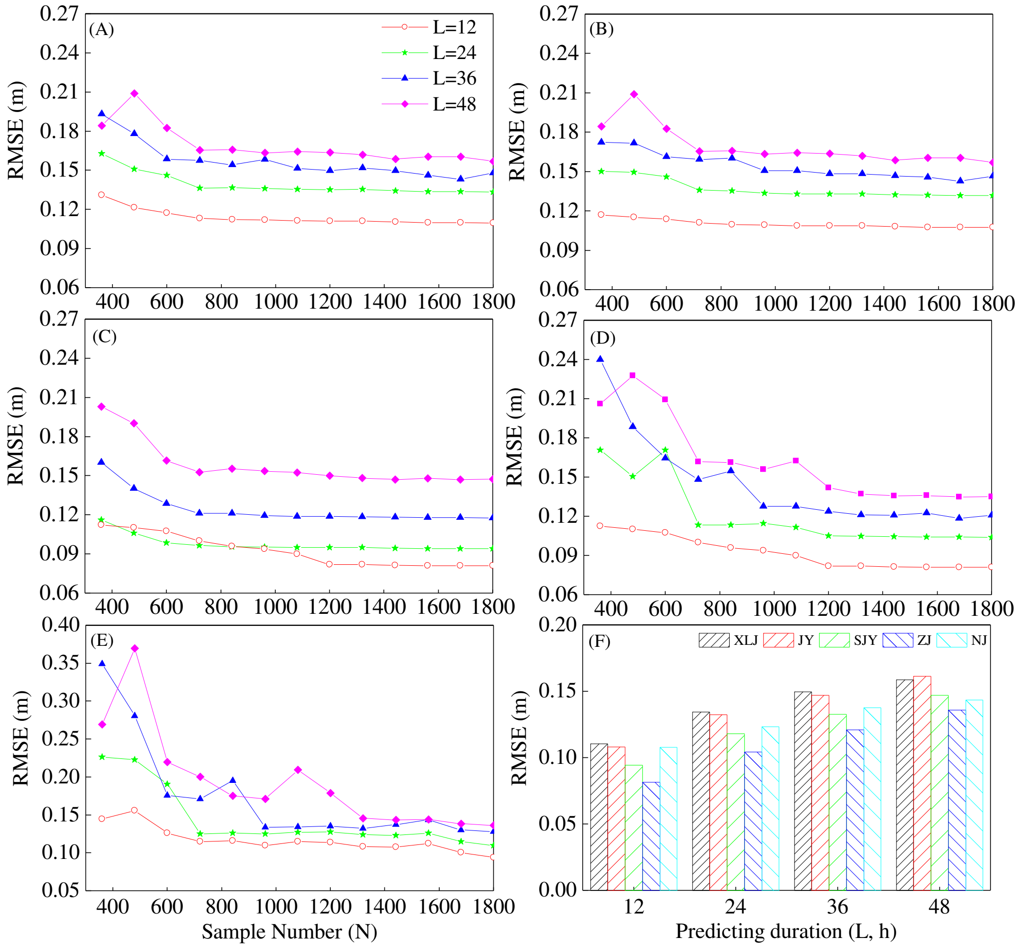

Fig. 9. Variation of the RSME values of the NS_TIDE&AR model with the number of samples (N) at: XLJ (A), JY (B), SJY (C), ZJ (D) and NJ (E) stations; and the comparison of the RMSE values when N=1440(F), for different predicting durations.

When the predictive errors of the NS_TIDE model are dynamically modelled by the AR model, the non-stationarity tests are conducted on the samples prior to constructing the AR model. It is found from the tests that the field data at this study mostly conforms the required stationarity. However, while larger temporal variations are discovered in the measurement data, the temporal gradients of the predictive errors are calculated and used in the AR analysis, and then an additional inverse transformation is used.

Following the construction of the AR model in Stage 1, it can now be applied as illustrated in **Fig. 7** (Stage 2). Taking the predictive errors of the NS_TIDE model in September 2016 (flood season) and January 2017 (dry season) at ZJ station as examples, **Fig. 10** shows the comparisons of the predictive errors from the NS_TIDE model and those estimated by the AR model. It can be clearly seen that the errors estimated by the AR analysis agree well with those from the NS_TIDE model. The errors in general are modulated by the tides and are larger during the flooding phase and lower during the ebbing phase. The envelope curve of the predictive errors of the NS_TIDE model in **Fig. 10** indicates the seasonal pattern for the predictive errors, which reflects its longer period variations. This can also be seen as the spectral peak at D_0 band in **Fig. 6**. The results clearly illustrate that the AR analysis is capable of estimating the predictive errors from the NS_TIDE model once it is calibrated and trained up a high level of accuracy, which provides an effective remedy for the NS_TIDE model in increasing its accuracy in water level predictions.

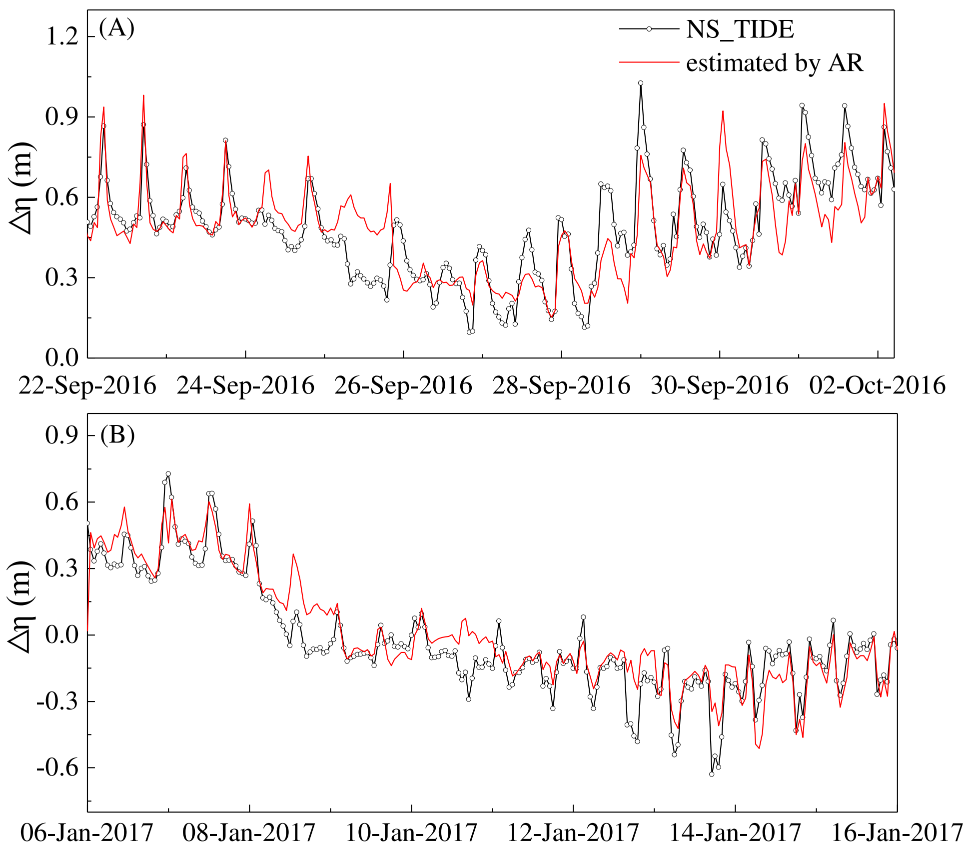

Fig. 10. Comparison of the predictive errors from the NS_TIDE model and those estimated by the AR analysis over flood (A) and dry (B) seasons at ZJ station

To examine the overall performance, the hybrid NS_TIDE&AR model is applied to the study site over the entire period between March 2015 and December 2017, where the measurements are available. For the 24-h (L=24) prediction, **Fig. 11(A)** compares the predicted water levels from both the NS_TIDE and the hybrid NS_TIDE&AR models with the observed water levels in 2016 at ZJ station and **Fig. 11(B)** shows the corresponding difference for the sake of clarity. The predicted water levels at ZJ station from both models agree well with the measurements in general, but their differences shown in **Fig. 11(B)** clearly indicate that the hybrid NS_TIDE&AR model outperforms the NS_TIDE model with a significant improvement. The predictive errors associated with the seasonal variation pattern from the NS_TIDE model are largely eliminated by the hybrid model, and therefore overall accuracy of the predictions is significantly improved.

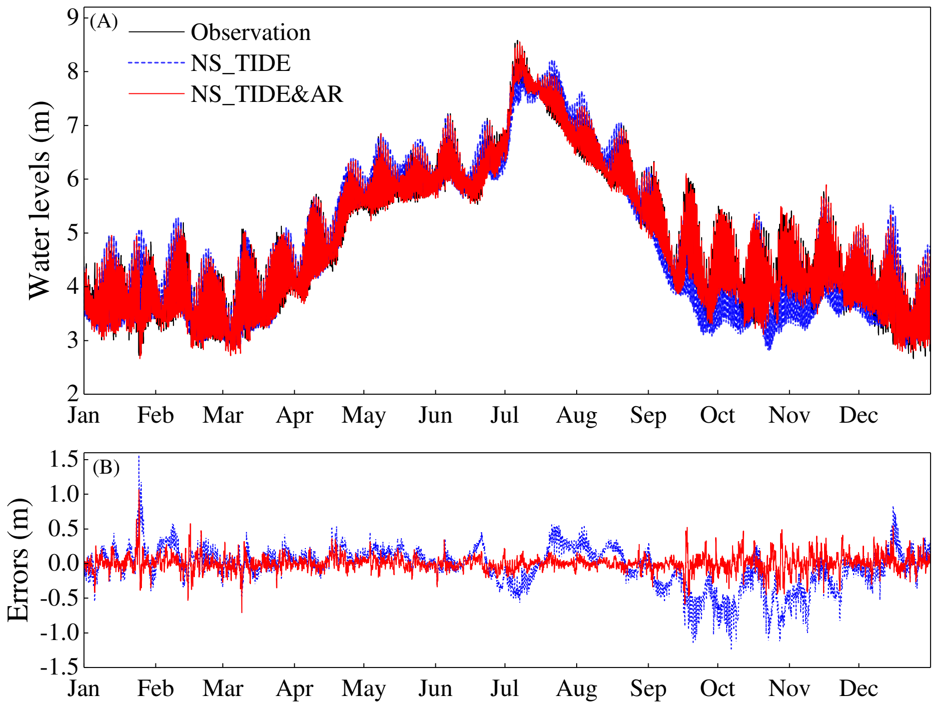

Fig. 11. Comparison of all the observed water levels in 2016 and the water levels predicted by the NS_TIDE and NS_TIDE&AR models (A) and their predictive errors (B) at ZJ station.

For the purpose of flood protection, water level prediction during the flood seasons is particularly important. Therefore, **Fig. 12** further shows the comparison of the predicted water levels by the NS_TIDE and NS_TIDE&AR models with the measurements at ZJ station during two flood seasons in 2016 and 2017. The predicted water levels from the NS_TIDE&AR model are found to match much better with the measurements than those from the NS_TIDE model in flood seasons, where the over- and under- predictions of the water level from the NS_TIDE model are largely corrected.

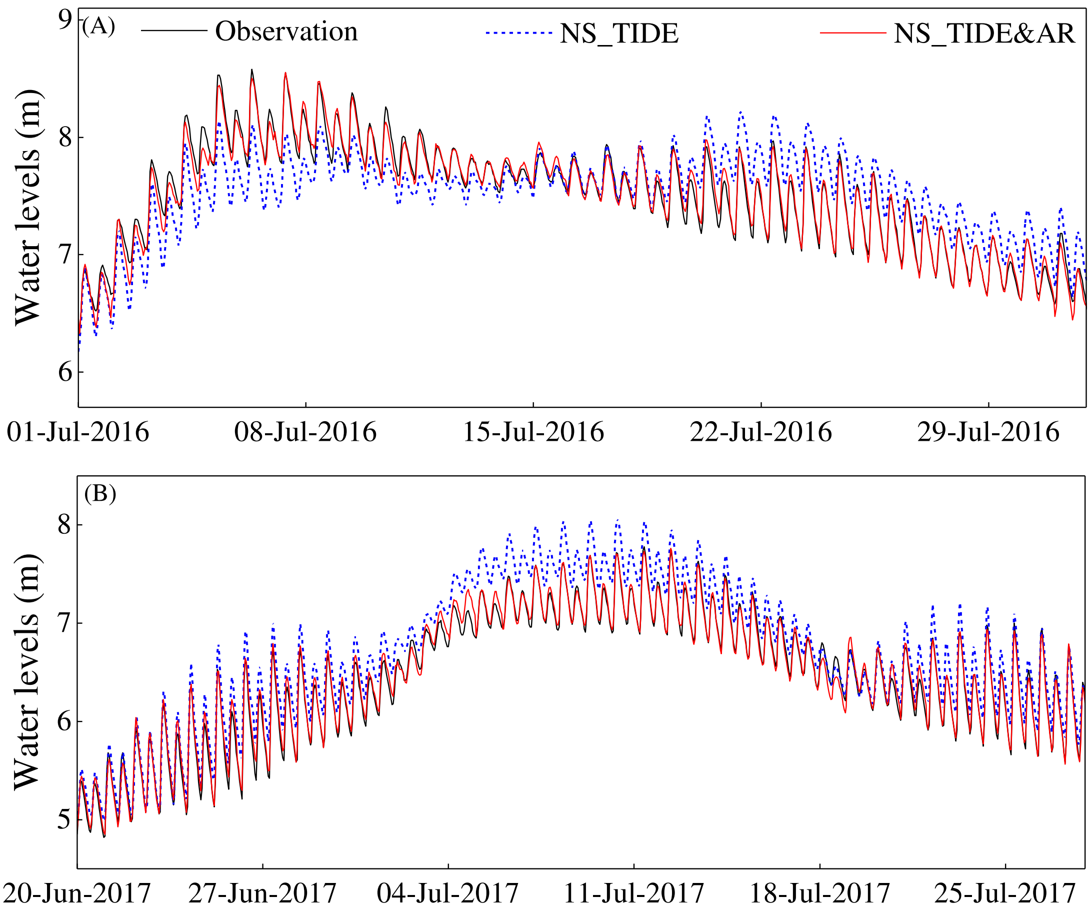

Fig. 12. Comparison of the water levels predicted by the NS_TIDE and NS_TIDE&AR models with the measurements at ZJ station during flood seasons of 2016 (A) and 2017 (B).

Fig. 13 shows the scatter plots of the predicted water levels from both the NS_TIDE and NS_TIDE&AR models against the measurements at all stations: XLJ, JY, SJY, ZJ and NJ for the 24-h prediction duration. The results clearly show that the predicted water levels from the NS_TIDE&AR model agree better with the measurements than those from the NS_TIDE model at all stations. At NJ and ZJ stations, the improvement of the NS_TIDE&AR model over the NS_TIDE model is similar for all waters, but at SJY, JY and XLJ stations, the improvement is seen progressively less significant particularly at the mean water level, as the tide forcing strengthens towards the estuary mouth as expected. Although, there are occasional outliers from the hybrid model from the perfect fit (45° line), the overall performance of predicting the water levels by the NS_TIDE&AR model at all 5 stations, nevertheless, is significantly improved in comparison with the NS_TIDE model.

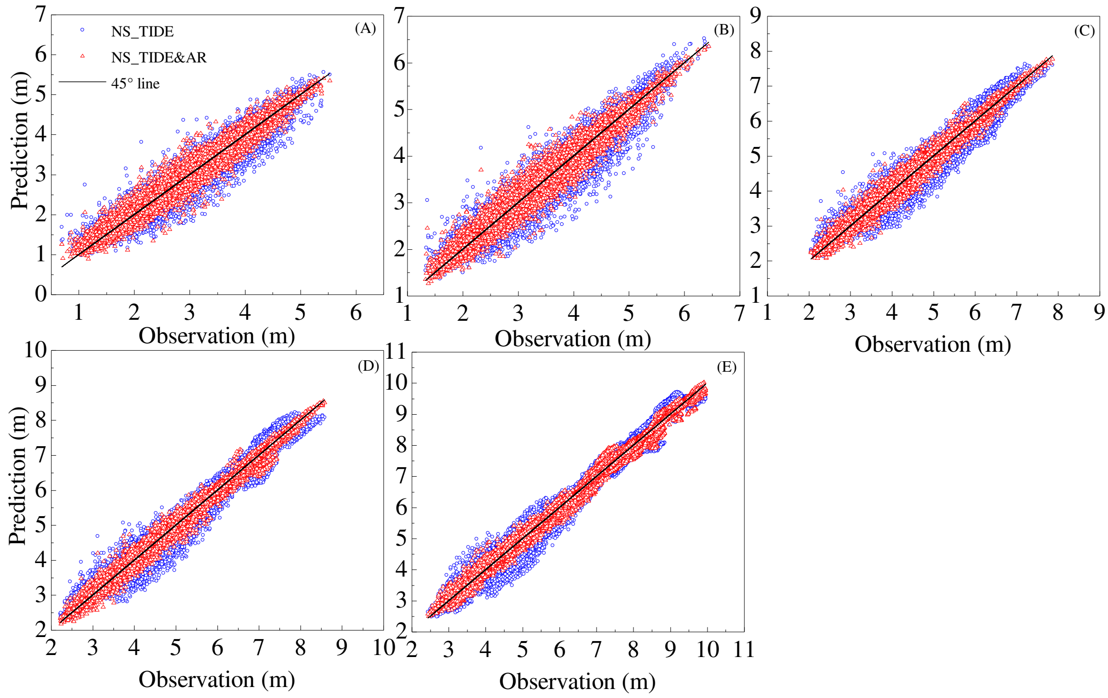

Fig. 13. Scatter plots of the water levels predicted by the NS_TIDE and NS_TIDE&AR models with the measurements at: XLJ (A), JY (B), SJY (C), ZJ (D) and NJ (E) stations.

6. DISCUSSION

To further understand the performance of the hybrid NS_TIDE&AR model, the hourly water levels are predicted by the hybrid model at all stations over a range of short-term prediction durations, namely 12, 24, 36 and 48 hours ahead. **Fig. 14(A)** compares the RMSE values of the predicted water levels from the hybrid NS_TIDE&AR and NS_TIDE models against the measurements respectively. Since the RMSE values of the predicted water levels by the NS_TIDE model for durations of 12, 24, 36, and 48 hours are almost the same, therefore, only one value is presented at each station in the figure. The results clearly show that the hybrid model significantly reduced the RMSE values in call cases. Taking the 24-hour prediction as an example, the hybrid model significantly reduces the RMSE values of the NS_TIDE model from 0.22 ~ 0.26 m to 0.10 ~ 0.13 m. Even for the 48-hour prediction, the longest in the tests, the RMSE values from the hybrid model are all less than 0.16 m.

Fig. 14(B) shows the RMSE values for the high tides, which are important when considering the flood protection in the estuary. The performance of the hybrid model is also much improved in comparison with the NS_TIDE model, although the overall RMSE values are slightly higher than those shown in **Fig. 14(A)**, but all within 0.02 m. For navigation, where accurate low water level predictions are as important as the high water levels, the RMSE values of the low water level predictions are presented in **Fig. 14(C)**. The predictive accuracy of the low water level prediction of the NS_TIDE&AR model is even better than that for the high water levels shown in **Fig. 14(B)**. In summary, the NS_TIDE&AR model significantly improves the predictive accuracy of the NS_TIDE model for both hourly water levels, high water levels, and low water levels.

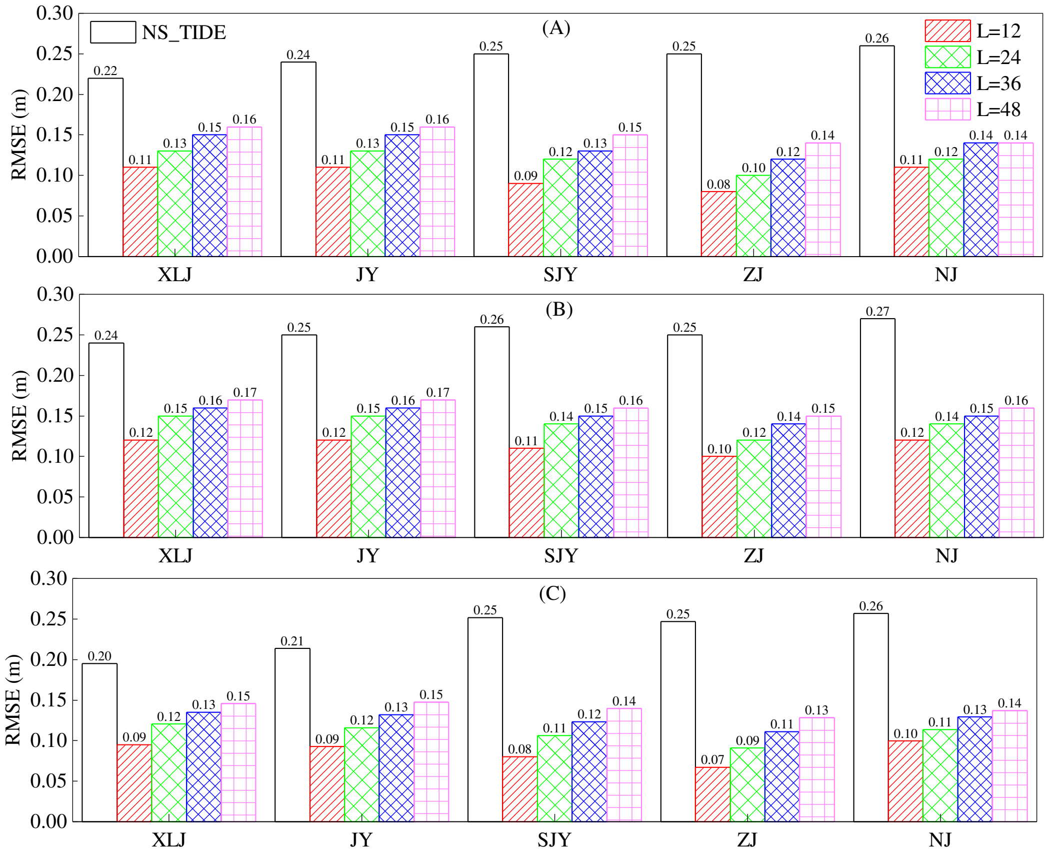

Fig. 14. RMSE values of the predicted water levels by the NS_TIDE and NS_TIDE&AR models over durations (L) of 12, 24, 36 and 48 hours for: all water levels (A); high water levels (B); and low water levels (C)

Spectral analysis is also conducted on the time series of the predictive difference of the hybrid model to the measurements in the frequency domain. **Fig. 15** shows the spectral energy density of the predictive errors of the hybrid NS_TIDE&AR model at all 5 stations in comparison with that of the NS_TIDE model as shown in **Fig. 6**. It can be seen that the peaks of the spectral energy density of the NS_TIDE&AR model are now much smaller in the D_0 frequency band than that of the NS_TIDE model, which means the NS_TIDE&AR model achieves a significant improvement over the NS_TIDE model in predicting the water level in the subtidal D_0 band. In the D_1, D_2 and D_4 tidal bands, the spectral peaks of the NS_TIDE&AR model are also found to be relatively smaller than those from the NS_TIDE model.

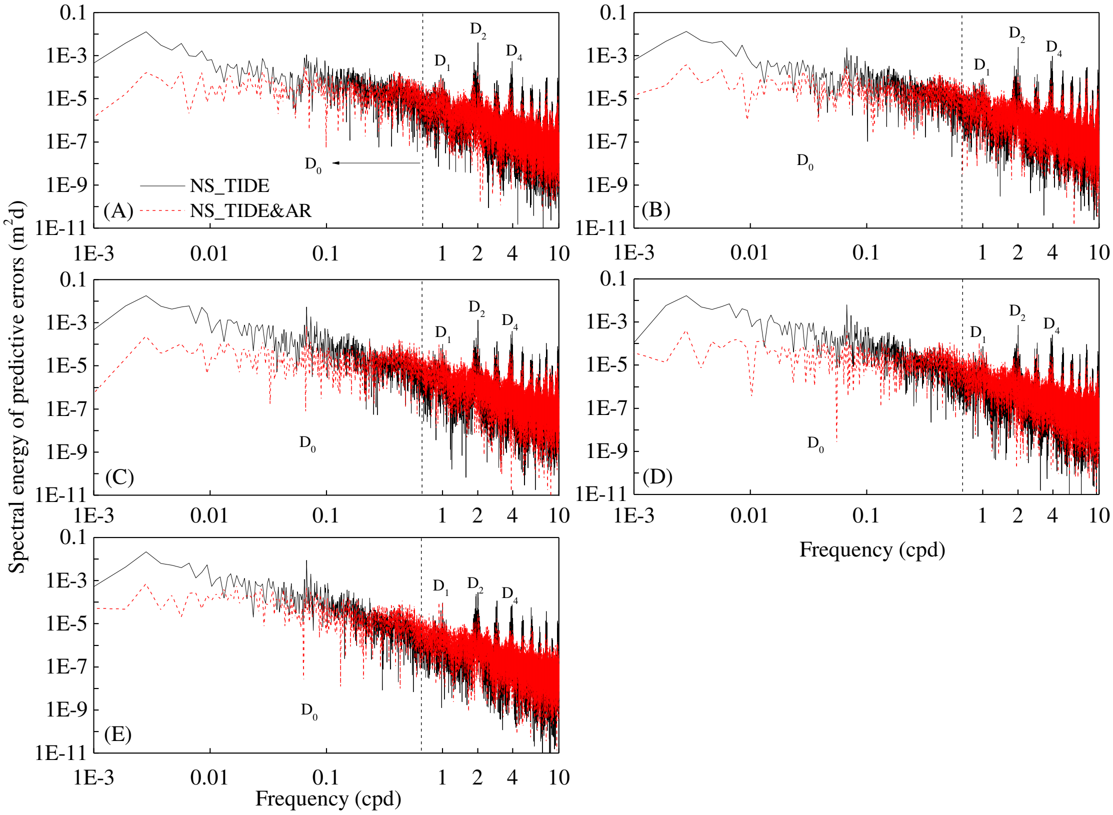

Fig. 15. Spectral energy distribution of the predictive errors of the NS_TIDE&AR and NS_TIDE models at: XLJ (A), JY (B), SJY (C), ZJ (D) and NJ (E) stations.

7. CONCLUSIONS

In this study, the auto-regressive (AR) analysis is developed and implemented with the existing NS_TIDE model for predicting the short-term water levels in the Yangtze estuary. The results show that by treating the predictive errors of the NS_TIDE model as a stochastic process, the AR analysis is capable of correlating robustly the predictive errors over short time periods (~48 hours), which can be used to effectively correct the predictions of water levels from the NS_TIDE model or possibly any other tide-prediction models. For the 24-h prediction, the RMSE values of the predicted water levels are less than 0.13 m, in comparison with those of 0.22 ~ 0.26 m of the NS_TIDE model without implementing the AR analysis. Spectral analysis indicates that the main improvement of the NS_TIDE&AR model over the NS_TIDE model is in describing the subtidal tides. Therefore, using AR analysis to estimate the predictive errors of the NS_TIDE model can significantly improve the short-term predictions of the water levels in estuary under the influence of strong river discharge, such as those in the Yangtze estuary.

CRediT authorship contribution statement

Yongping Chen: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Funding acquisition, Supervision. **Min Gan:** Conceptualization, Methodology, Software, Writing - Original Draft, Writing - Review & Editing, Validation, Formal analysis, Investigation, Funding acquisition. **Shunqi Pan**: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Funding acquisition, Supervision. **Haidong Pan**: Resources, Investigation. **Xian Zhu**: Data Curation, Investigation. **Zhengjin Tao**: Data Curation, Visualization.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partly supported by the National Key R&D Program of China [Grant No: 2017YFC0405401], the Fundamental Research Funds for the Central Universities of China [Grant Nos: 2017B20214 and 2018B635X14] and the Postgraduate Research & Practice Innovation

Program of Jiangsu Province [Grant No: KYCX18_0602]. The second author also would like to acknowledge the financial support from the China Scholarship Council (CSC) under PhD exchange program [201906710022] with Cardiff University. The authors would also like to thank Pascal Matte for providing the NS_TIDE model software package.

References

Cai, H., Savenije, H. H. G., & Toffolon, M. (2014). Linking the river to the estuary: influence of river discharge on tidal damping. *Hydrol. Earth Syst. Sci.*, 18, 287-304, doi: 10.5194/hess-18-287-2014.

Carbajal-Hernández, J.J., Sánchez-Fernández, L.P., Carrasco-Ochoa, J.A. & Martínez-Trinidad, J.F. (2012). Assessment and prediction of air quality using fuzzy logic and autoregressive models. *Atmospheric Environment*, 60, 37-50, doi:10.1016/j.atmosenv.2012.06.004.

Cavaliere, G. & Georgiev, I. (2007). Testing for Unit Roots in Autoregressions with Multiple Level Shifts. *Econometric Theory*, 23, 1162-1215, doi: 10.1017/S0266466607070466

Chen, J.C. & Boccelli, D.L. (2018). Forecasting hourly water demands with seasonal autoregressive models for real-time application. Water Resources Research, 54, 879–894, doi: 2017WR022007.

Chen, W., Chen, K., Kuang, C.P., Zhu, D.Z., He, L.L., Mao, X.D., Liang, H.D., & Song, H.L. (2016). Influence of sea level rise on saline water intrusion in the Yangtze River Estuary, China. *Applied Ocean Research*, 54, 12-25, doi:10.1016/j.apor.2015.11.002.

Codiga, D.L. (2011). Unified tidal analysis and prediction using the UTide Matlab functions, Technical Report 2011-01. Graduate School of Oceanography, University of Rhode Island, Narragansett, RI. 59pp. [Available online at ftp://www.po.gso.uri.edu/pub/downloads/codiga/pubs/2011Codiga-UTide-Report.pdf.]

Gallo, M.N. & Vinzon, S.B. (2005). Generation of overtides and compound tides in Amazon estuary. *Ocean Dynamics*, 55, 441-448, 10.1007/s10236-005-0003-8.

Gan, M., Chen, Y., Pan, S., Li, J., & Zhou, Z. (2019). A modified nonstationary tidal harmonic analysis model for the Yangtze estuarine tides. *J. Atmos. Ocean. Technol.*, 36, 513-525, doi:10.1175/JTECH-D-18-0199.1.

Ge, M. & Kerrigan, E.C. (2016). Short-term Ocean Wave Forecasting Using an Autoregressive Moving Average Model. *11th International Conference on Control (CONTROL)*, Belfast, UK, doi: 10.1109/CONTROL.2016.7737594.

Guo, L.C., Wegen, M.V. D., Jay, D.A., Matte, P., Wang, Z.B., Roelvink, D., & He, Q. (2015). River-tide dynamics: Exploration of nonstationary and nonlinear tidal behavior in the Yangtze River estuary. *J. Geophys. Res. Oceans*, 120, 3499-3521, 10.1002/2014JC010491.

Guo, L.C., Wegen, M.V. D., Wang, Z.B., Roelvink, D., & He, Q. (2016). Exploring the impacts of multiple tidal constituents and varying river flow on long-term, large-scale estuarine morphodynamics by means of a 1-D model. *J. Geophys. Res. Oceans*, 121(5):1000-1022, 10.1002/2016JF003821.

Holland, P. W. & Welsch, R.E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics-Theory and Methods*, 6(9), 813-827, doi:10.1080/03610927708827533.

Jay, D. A. (1991). Green’s law revisited: Tidal long-wave propagation in channels with strong topography. *J. Geophys. Res.*, 96(C11), 20 585–20 598, doi: 10.1029/91JC01633.

Jay, D. A., Leffler, K., & Degens, S. (2011). Long-Term Evolution of Columbia River Tides. *J.*

Waterw. Port Coastal Ocean Eng., 137, 182-191, doi:10.1061/(ASCE)WW1943-5460.0000082.

Kukulka, T. & Jay, D.A. (2003a). Impacts of Columbia River discharge on salmonid habitat: 1. A nonstationary fluvial tide model. *J. Geophys. Res. Ocean*, 108, 3293, doi:10.1029/2002JC001382.

Kukulka, T. & Jay, D. A. (2003b). Impacts of Columbia River discharge on salmonid habitat: 2. Changes in shallow-water habitat. *J. Geophys. Res. Ocean.*, 108, 3294, doi: 10.1029/2003JC001829

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. How Sure Are We That Economic Time Series Have Unit Root? *Journal of Econometrics*, 54, 159-178, doi: 10.1016/0304-4076(92)90104-Y.

Leffler, K.E. & Jay, D. A. (2009). Enhancing tidal harmonic analysis: Robust (hybrid L1/L2) solutions. *Cont. Shelf Res.*, 29, 78-88, doi:10.1016/j.csr.2008.04.011.

Li, M., Wang, J.Q., Bennett, J.C., & Robertson, D. E. (2015). A strategy to overcome adverse effects of autoregressive updating of streamflow forecasts. 19, 1–15, doi: 10.5194/hess-19-1-2015.

Lu, S., Tong, C.F., Lee, D., Zheng, J.H., Shen, J., Zhang, W., & Yan, Y.X. (2015). Propagation of tidal waves up in Yangtze Estuary during the dry season. *J. Geophys. Res. Ocean.*, 120, 6445-6473, doi:10.1002/2014JC010414.

Mandal, S., Witz, J.A., & Lyons, G.J. (1992). Reduced order ARMA spectral estimation of ocean waves. *Applied Ocean Research*, 14, 303-312, doi: 10.1016/0141-1187(92)90034-H.

Matte, P., Jay, D.A., & Zaron, E.D. (2013). Adaptation of Classical Tidal Harmonic Analysis to Nonstationary Tides, with Application to River Tides. *J. Atmos. Ocean. Technol.*, 30, 569-589, doi:10.1175/JTECH-D-12-00016.1.

Matte, P., Secretan, Y., & Morin, J. (2014). Temporal and spatial variability of tidal-fluvial dynamics in the St. Lawrence fluvial estuary: An application of nonstationary tidal harmonic analysis. *J. Geophys. Res. Oceans*, 119, 5724-5744, doi:10.1002/2014JC009791.

Minguez, R., Abascal, A.J., Castanedo, S., & Medina, R. (2012). Stochastic Lagrangian trajectory model for drifting objects in the ocean. *Stoch Environ Res Risk Assess*, 26:1081-1093, doi: 10.1007/s00477-011-0548-7.

Mirzavand, M. & Ghazavi, R. (2015). A stochastic modelling technique for groundwater level forecasting in an arid environment using time series methods. *Water Resources Management*, 29, 1315-1328, doi:10.1007/s11269-014-0875-9.

Pan, H.D., Guo, Z., Wang, Y.Y., & Lv, X.Q. (2018). Application of the EMD method to river tides. *J. Atmos. Ocean. Technol.*, 35, 809–819, doi:10.1175/JTECH-D-17-0185.1.

Pawlowicz, R., Beardsley, B., & Lentz, S. (2002). Classical tidal harmonic analysis including error estimates in MATLAB using T_TIDE. *Comput. Geosci.*, 28, 929-937, doi:10.1016/S0098-3004(02)00013-4.

Petaccia, P., Serravall, R., & Pellicano, F. (2006). Improved method of sea level forecasting at Venice (Northern Adriatic Sea). *Communications in Nonlinear Science and Numerical Simulation*, 11, 281-296, doi: 10.1016/j.cnsns.2004.11.008.

Peters, T.C., Kar, T.R., Jamas, P.F., Knight, R., & Easterlin, D. (1998). First difference method: Maximizing station density for the calculation of long-term global temperature change. *J.*

Geophys. Res. Atmospheres., 103, 967-974, doi: 10.1029/98JD01168.

Savenije, H.H.G. (2015). Prediction in ungauged estuaries: An integrated theory. *Water Resour. Res.*, 51, 2464-2476, doi:10.1002/2015WR016936.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, 63, 117-126, doi: 10.2307/2335091.

Torres, J.L., García, A., Blas, M. D., & Francisco, A. D. (2005). Forecast of hourly average wind speed with ARMA models in Navarre (Spain). *Solar Energy*, 79, 65–77, doi: 10.1016/j.solener.2004.09.013.

Turki, I., Laignel, B., Kakeh, N., Chevalier, L., & Costa, S. (2015). A new hybrid model for filling gaps and forecast in sea level: application to the eastern English Channel and the North Atlantic Sea (western France). *Ocean Dynamics*, 65(4), 509-521, doi: 10.1007/s10236-015-0824-z.

Zhang, E. F., Savenije, H. H. G., Chen, S.L., & Mao, X.H. (2012). An analytical solution for tidal dynamics in the Yangtze Estuary, China. *Hydrol. Earth Syst. Sci. Discuss.*, 9, 2213–2244, doi: 10.5194/hessd-9-2213-2012.