International Conference on Hydroinformatics

2014

Comparison Of Three Methods For Spatial Distribution Of Error-Correction Algorithms

Xuan Wang

Vladan Babovic

COMPARISON OF THREE METHODS FOR SPATIAL DISTRIBUTION OF ERROR-CORRECTION ALGORITHMS

XUAN WANG (1), VLADAN BABOVIC (1, 2)

(1): Department of Civil and Environmental Engineering, EW1-01-14, Engineering Drive 2, National University of Singapore, Singapore 117576

(2): NUSDeltares, National University of Singapore, E1-08-24, Engineering Drive 2, Singapore 117576

Contract/grant sponsor: Singapore-Delft Water Alliance (SDWA); contract/grant number: R-264-001-001-272

ABSTRACT

Data assimilation is a useful tool to correct the discrepancies of numerical model results by extracting reliable information from observed data. One of popular data assimilation techniques is the spatial distribution based on error-correction, since it can address the challenge when number of monitoring stations is limited. Current research only focuses on the estimation of spatial distribution pattern, or the improvement of the competence of different spatial distribution methods, but lacks the comparison either in their characteristics or in the performances. In this study, we compared three different approaches, Kriging, Artificial Neural Network (ANN) and inter-model correlation inspired by Kalman Gain, for spatial distribution on error correction. Based on the application in a real case of Singapore Regional model, the performance and adaptive capabilities of these methods are analyzed through testing the sensitivity in response to different observation points and hydrodynamic regimes. The results suggest that the performance varies among different methods and changes with various scenarios, indicating that an appropriate selection of algorithms under different environmental condition is necessary.

Key words: Data assimilation, Error correction, Spatial distribution, Kriging, inter-model, ANN

INTRODUCION

Numerical modeling is one of the most popular means to simulate and forecast the state of oceanographic systems. However, such kind model tends to produce imperfect results due to several reasons, such as model resolution, parameter uncertainty, simplifying assumptions, absence of data for proper setting of boundary and initial conditions. Data assimilation, which combines the results from numerical model with the measurements, can help combat the inevitable presence of model error and hence allow a numerical model to approximate the actual sea condition more closely [1-3].

Kalman Filter (KF) [4,5] is a widely-practiced data assimilation approach. It has been applied in several oceanographic and meteorological applications [6,7]. However, one of its major drawbacks is that it requires huge computational resource associated with the error propagation. Besides, it is also limited to a forecasting horizon where the improved initial conditions are washed out [8].

Another data assimilation technique is model error correction. This method corrects the output variables of the model directly, and hence can be executed offline to the numerical model [9]. Normally, it is carried out based on two steps, error forecasting at measured locations and then error distribution to all the other locations without measurements. This paper only focuses on the latter step, i.e. model error correction through spatial distribution. For this area, most research only focuses on the estimation of spatial distribution pattern, or the improvement of the competence of different spatial distribution methods, but lacks the comparison either in their characteristics or in the performances. In this study, we compared three different approaches, Kriging, inter-model correlation inspired from Kalman Gain and Artificial Neural Network (ANN), for spatial distribution on error correction.

To examine the performance of the above methods, these methods are applied in a real case of Singapore Regional model (SRM) to correct the water level outputs directly.

ALGORITHM

The numerical model error, or residual, is defined as the difference between the actual measurements and numerical model results.

$$\varepsilon = x^{mea} - x^{num} \tag{1}$$

Where, ε the model residual, the x^{mea} the measurements, x^{num} the numerical model output;

Given a group of known model residual ε_o at the observed sites, the residual at the non-measured sites ε_u can be estimated through the technique of spatial distribution.

Residual Distribution with Approximated Ordinary Kriging

Ordinary Kriging is one of the most popular spatial interpolation techniques and it is also applied in the area of environment engineering [10]. The fundamentals of the algorithm estimate the value at a non-measured points s_p based on a series of observed values $(z_i = z(s_i), i = 1,..n)$ at nearby measured points s_i. A Kriging estimator $\hat{z}_p$ is a linear combination of z_i which can be expressed as follows:

$$\hat{z}_p = \sum_{i=1}^{n} w_{pi} z_i \tag{2}$$

In equation (2), z_i denotes the values at a nearby measured point s_i, w_{pi} the weight between s_i and non-measured point s_p, and n the number of nearby measured points. The weight w_{pi} is calculated according to the variogram.

The variogram involves both experimental variogram and model variogram. The experimental variogram γ_{ij} (also referred to as sample variogram) at the measured points is estimated from the observations at sampling points. And the variogram involving non-measured points (i.e. model variogram) γ_{ip} is then computed using base functions of a certain class (e.g. linear model, exponential model, Gaussian model, or spherical model). However, choosing appropriate variogram base functions and fitting them to data remain among the most controversial topics in Kriging methods [11]. Therefore Wang and Babovic [12] suggested the "Approximated Ordinary Kriging" method. It expresses the spatial dependence structure via an approximated variogram which approximates the spatial relationship of the observed phenomenon. It can be calculated and applied in the following procedure:

(a) Estimate approximated variogram $\hat{\gamma}_{ij}$:

If the variogram is only dependent on the length of distance but not its direction, it can be estimated as

$$\hat{\gamma}_{ij} = \gamma^{y}(h_{ij}) = \gamma^{y}_{ij}(s_i + h_{ij}, s_i) = \frac{1}{2N}\sum_{t=1}^{N}(y_t(s_j) - y_t(s_i))^2 \tag{3}$$

where, $\hat{\gamma}_{ij}$ denotes approximated variogram and $y_t(s_i), y_t(s_j)$ the value of variable x^{num} at location s_i and s_j.

(b) Calculate weights w_{pi}

The weights w_{pi} are computed from the Kriging linear equations:

$$\begin{cases} \sum_{i=1}^{n} w_{pi}\gamma_{ij}(s_i, s_j) + \mu = \gamma_{pj}(s_j, s_p) \qquad (j = 1,...,n) \\ \sum_{i=1}^{n} w_{pi} = 1 \end{cases} \tag{4}$$

where, γ_{pj} is the value of variogram between measured point s_j and non-measured point s_p; μ is the Lagrange multiplier.

(c) Estimate variable $\hat{\varepsilon}(s_p)$

The residual at non-measured points $\hat{\varepsilon}_t(s_p)$ is interpolated through equation (2) with weights calculated in step (b) above based on the data at nearby measured location.

Residual Distribution with Inter-Model Correlation

Drawn inspiration from Kalman filter, Mancarella *et al.* [13] suggested building an error distribution scheme based on inter-model relationship. And the model residual can be distributed through a linear inter-model structure based on numerical model output.

With the residues forecasted at measured locations, the corrected model output in a set of non-measured locations is given by:

$$x_u^c = x_u^{num} + \hat{\varepsilon}_o \times W_{ou} \tag{5}$$

Where x_u^c the corrected numerical model output; x_u^{num} the numerical model output; the subscript *u* indicates non-measured locations and *o* the observed locations; $\hat{\varepsilon}_o$ is the residues forecasted at observed locations; W_{ou} is the linear model created to describe relationships between observed locations and non-measured locations.

It is obtained by

$$x_o^{num} \times W_{ou} = x_u^{num} \tag{6}$$

Where x_o^{num} the data matrix of the numerical model output at observed locations.

Residual Distribution with ANN

The first-order approximation from the above inter-model is a simple and fast distribution scheme. However, in consideration of limitation of the linear spatial weights, Wang et.al [14] suggested to apply Artificial Neural Network (ANN) to establish the non-linear nature of spatial distribution of residues in ocean hydrodynamic simulations.

The spatial weighting function is estimated with ANN, based on the numerical model output, to approximate the spatial relationship between locations. The procedure is carried out in the following three steps:

Step 1: evaluate the spatial weighting function Ŵ with ANN

The structure used by ANN to estimate the weighting function is constructed and indicated in Figure 1. The numerical model outputs $x_o^{num}(s_i)$ at the observed locations S_i are used as input for the ANN structure, and the model output $x_u^{num}(s_p)$ at the non-measured locations S_p are utilized as target output to train the structure of ANN.

Step 2: assess the model error at non-measured locations $\hat{\varepsilon}(s_p)$

After calculating the model error $\hat{\varepsilon}_o(s_i)$ at observed locations, the error $\hat{\varepsilon}_u$ at non-measured locations can be assessed by the ANN structure trained before, with the $\hat{\varepsilon}_o(s_i)$ being the input of the weighting function i.e. $\hat{\varepsilon}_u = \hat{W}(\hat{\varepsilon}_o)$.

Step 3: correct the numerical model output at non-measured locations

The numerical model output x_u^{num} at the non-measured locations can be corrected as follow:

$$x_u^c = x_u^{num} + \hat{\varepsilon}_u \quad (7)$$

Where, Ŵ is the weighting function trained by ANN.

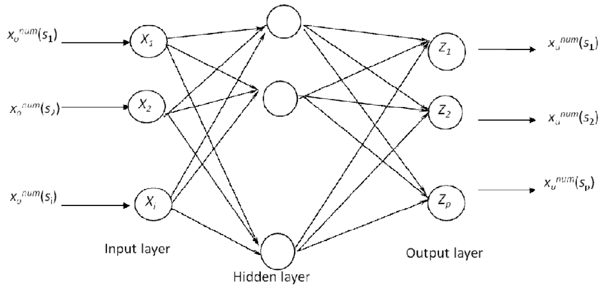

Figure 1: Architectural structure of spatial weights estimation

THE HYDRODYNAMIC MODEL

The Singapore Regional Model (SRM) is implemented within Delft3D Flow system to provide hydrodynamic information, in particular sea level anomalies (residual water levels) in the Singapore Straits [15]. The model was set up with 3 open boundaries, which are the South China Sea on the east, the Andaman Sea on the West and a small part of Java Sea on the South. Along the open boundaries water level variation is prescribed by 8 tidal constituents (Q1, O1, P1, K1, N2, M2, S2 and K2).

However, for computational efficiency a 3×3 aggregated coarse grid version of the SRM (also abbreviated as "SRMC") was built up with 4239 cells. It has been tested in Wang [12] that, such aggregated model is eligible to provide background information for the error correction scheme and thus used in this study. The scope, grid and bathymetry of SRMC are shown as Figure 2.

The simulation was carried out from 2004 Jan. 1st 00:00 to 2004 Dec. 31st 00:00 with a time step of 4 mins and hourly recording. It produced 8761 hourly time series of water level for all grid points in the domain. In order to eliminate the influence of the initial condition, the first 10 days of data points were discarded.

Thirteen stations are considered in the present study. West Coast, Tanjong Changi, Tanah Merah, Sembawang and Raffles are located around Singapore Region, Langkawi, Kelang, Lumut and Penang are located at the Malacca Strait, and Tioman, Getting, Kuantan as well as Sedili are located in the east of Malaysia peninsular. Their locations are shown in Figure 2. The measurements of water level in 2004 are available at these stations. Two stations are selected from each region (West Coast, Tanjong Changi, Langkawi, Kelang, Tioman and Getting) as measured stations, and the others are assumed to be non-measured ones. Four cases are tested in this study:

Case 1(the Singapore Region): correct results at Tanah Merah, Sembawang and Raffles based only on West Coast, Tanjong Changi;

Case 2 (the Malacca Strait): correct results at Lumut and Penang based on Langkawi and Kelang;

Case 3 (the east of Malaysia peninsular): correct results at Kuantan and Sedili based on Tioman, Getting;

Case 4 (the entire domain): correct results at the seven non-measured locations together based on the data at all the other six measured locations.

The model residuals at measured locations are forecasted off-line based on the linear local model (LM) applied by Babovic [16], which are not repeated in this study. The corresponding forecasting results are further utilized in above three distribution scheme. The first half year of 2004 are considered to be pre-operational period to set up the model and the second half year of 2004 is assumed as operational period. And the measurements here are only used for validation.

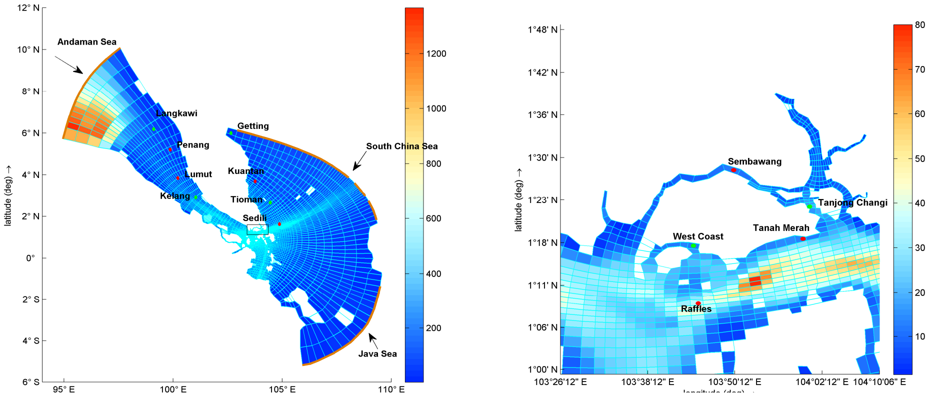

Figure 2. scope grid and bathymetry of coarse Singapore Regional Model, and the sample stations (green indicate the observed stations and red the assumed non-observed stations)

RESULTS AND DISCUSSION

The mentioned scheme distributed the forecasted model error with forecasting horizon T=1hour from measured stations to other points to correct the numerical model. The results are assessed in terms of root mean square error (RMSE) and percentage of improvement (imp%), as defined in Equ. (8-9).

$$RMSE = \sqrt{\frac{\sum (x^{mea} - x^{c})^2}{n}} \tag{8}$$

$$imp\% = \frac{RMSE_{SRMC} - RMSE}{RMSE_{SRMC}} \times 100\% \tag{9}$$

where, x^{mea} is the observed water level ; x^{c} is the water level after correction; and n is the number of records; $RMSE_{SRMC}$ is the root mean square error of original numerical model.

Take the correction results at forecasting horizon of 1hour as example, the comparison of results for the three different methods are shown in Figure 3, where the percentage of improvement at the seven non measured locations based on case 1, 2 and 3 are plotted. It can be seen from these two figures that the method of AOK can correct the numerical model by more than 50% for most stations. Both ANN and inter-model gain is also able to achieve comparable improvement at the stations in the area around Singapore. However, for other areas, the AOK shows significant advantage compared with the other two methods. It suggested that, compared with the linear gain matrix by inter model method and the non-linear spatial weighting function by ANN, the approximated variogram used by AOK can not only capture the spatial relationship more accurately but also be more adaptive in area with complex hydrodynamic condition. For the case 1,2, and 3 shown in Figure 3, the numerical model can be improved by 60% to 85% in the area around Singapore, which is higher than the Malacca Strait (case 2) and in the area of east of Malaysia peninsular (case 3). It means that all the three methods perform adequately for case 1. However, in the area of Malacca Strait, the performances of all the three methods deteriorate seriously and even the improvement through method of AOK decrease to 30%. In order to understand the spatial correlation between different locations in the whole study area, the coefficient correlations estimated by the actual observed water level is shown in Table 1. It can be seen that all the five stations in the area of Singapore Region have strong correlation with each other, all of which are higher than 0.90. Followed are the stations in the area of east of Malaysia peninsular. Although Getting seems weakly related to Sedili, the station of Tioman shows high correlation with both Kuantan and Sedili. However, in the area of Malacca Strait, all the four stations (Langkawi, Kelang, Lumut and Penang) have very weak correlation each other which are lower than 0.84. Therefore, one possible reason to explain the less effective performance in this area may be that the hydrodynamic condition at the two observed location (Langkawi and Kelang) do not have strong correlation to the other two non-observed locations (Lumut and Penang). The results in case 4 also show similar trends for the three areas. Another reason may be the hydrodynamic condition in the Malacca Strait and the simulation original numerical model is also challenged in this area. It may be because this area may be influenced by the Indian Ocean which has not be considered in the current numerical model.

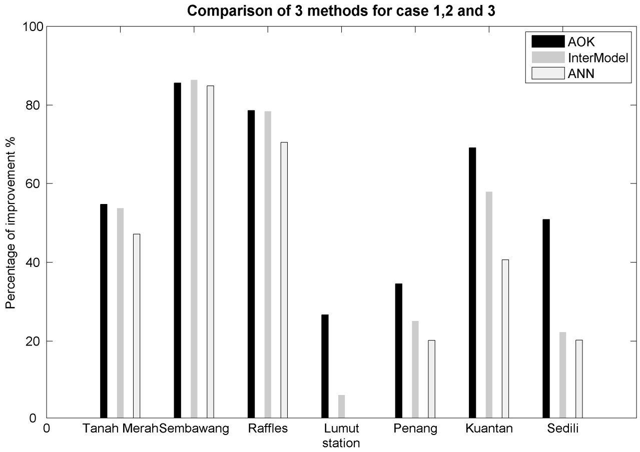

Figure 3: the comparison of percentage of improvement through AOK, inter-model and ANN for case 1,2 and 3

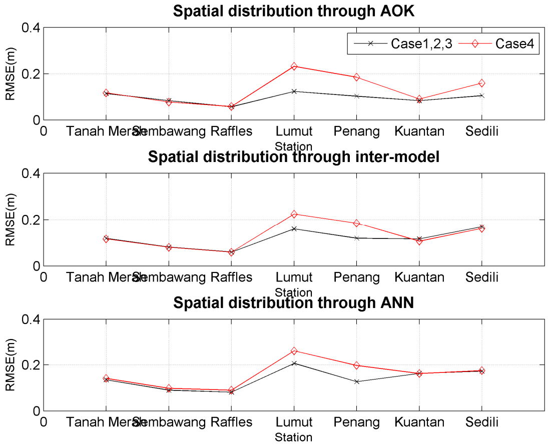

Figure 4: the comparison of RMSE after distribution in case4 and case 1,2,3

The RMSE of the correction result in case 4 is compared with that in case1, 2, or 3 through method AOK , ANN and inter-model gain are also shown in Figure 4. It can be seen that in the area around Singapore and the area of Malaysia peninsular, the tests of applying local spatial distributions separately in case 1 and 3 have similar RMSE with that of applying distribution within the entire area in case 4. However, for the area of Malacca Strait, the locally spatial distribution even produces less error than the global distribution. It suggests that including more locations as measurement input does not necessarily improve the distribution accuracy. Therefore, selecting the correlated observation location is more important than the amount of observation station to improve the efficacy of spatial distribution.

In order to further indicate distribution results directly, the station of Raffles and Kuantan are selected as example, the distributed model residuals and the water levels after correction through AOK are shown in Figure 5 and 6. It can be seen that the AOK method is capable to correct the water level from SRMC. It can capture their rising and falling tendencies with less error left.

*Table 1. The correlation coefficient of selected sample stations**

	wc	cf	lk	pn	tma	gt	tm	sb	rf	lm	pn	kt	sd
wc	1.000	0.945	0.767	0.884	0.268	0.075	0.937	0.951	0.998	0.724	0.447	0.078	0.552
cf		1.000	0.708	0.798	0.504	0.116	0.980	0.997	0.925	0.690	0.383	0.322	0.764
lk			1.000	0.909	0.229	0.035	0.674	0.736	0.775	0.440	0.838	0.126	0.398
pn				1.000	0.206	0.035	0.780	0.819	0.892	0.700	0.616	0.059	0.423
tma					1.000	0.766	0.486	0.469	0.220	0.446	0.079	0.975	0.933
gt						1.000	0.096	0.087	0.107	0.130	0.112	0.844	0.603
tm							1.000	0.974	0.917	0.702	0.339	0.303	0.742
sb								1.000	0.933	0.679	0.422	0.286	0.734
rf									1.000	0.716	0.464	0.031	0.505
lm										1.000	0.095	0.305	0.566
pn											1.000	0.121	0.061
kt												1.000	0.840
sd													1.000

*wc- West Coast, cf- Tanjong Changi, lk- lankawi, kl- Kelang, tma- Tioman, gt- Getting, sb- Sembawang, rf- Raffles, lm- Lumut, pn- Penang, kt- Kuantan, sd- sedili.

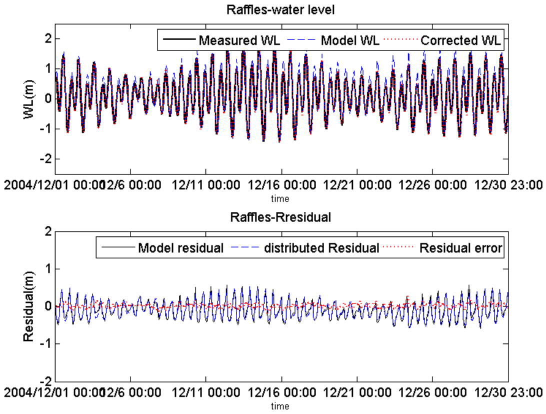

Figure 6: the corrected water level and distributed residual at station of Raffles

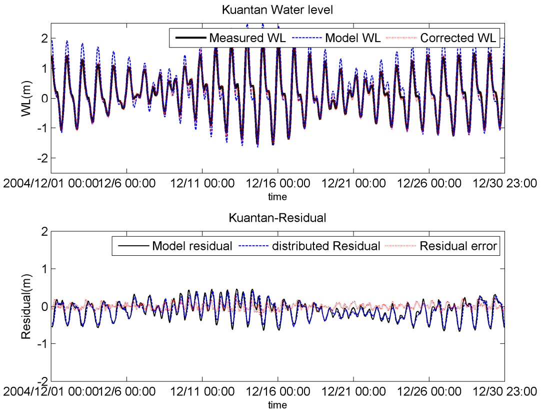

Figure 7: the corrected water level and distributed residual at station of Kuantan

CONCLUSION

Given the limitations of numerical modeling, the data assimilation method has become popular to further correct the numerical models. As part of these techniques, the limited measured location necessitates the spatial distribution technique to distribute the information from the location with observation to other non-observed location of interest. This paper discusses and compares three different spatial distribution methods, the approximated Ordinary Kriging (ANN) inspired by the Ordinary Kriging, the Artificial Neuron Network (ANN) and the inter-model gain inspired by the Kalman gain. The results show that for the area (e.g. in the area of Singapore region) where the selected locations have strong correlation each other, all these three methods perform adequately and can remove the error effectively. However, for the area where the hydrodynamic condition is complex and the correlation of selected stations is not strong enough (e.g. the Malacca Strait), only the method AOK is able to correct the numerical model with 30% error removed although the improvement is lower than it has done in the other area. The finding indicates that compared with the linear gain matrix by inter model method and the non-linear spatial weighting function by ANN, the approximated variogram used by AOK can not only capture the spatial relationship more accurately but also be more adaptive in area with complex hydrodynamic condition. In addition, through the four case tests, we compare the performance of local and global spatial distribution. The results suggest that including more locations as measurement input does not necessarily improve the distribution accuracy. Therefore, selecting the correlated observation location is more important than the amount of observation station to improve the efficacy of spatial distribution.

REFERENCE

[1] Babovic, V., Canizares, R., Jensen, H.R., Klinting, A., 2001. Neural networks as routine for error updating of numerical models. *Journal of Hydraulic Engineering*, **127**(Compendex): 181-193.

[2] Vojinovic, Z., Kecman, V., 2003. Data assimilation using recurrent radial basis function neural network model, CIMSA'03. 2003 *IEEE International Symposium on Computational Intelligence for Measurement Systems and Applications*, 29-31 July 2003. IEEE, Piscataway, NJ, USA, pp. 61-6.

[3] van den Boogaard, H., Mynett, A., 2004. Dynamic neural networks with data assimilation. *Hydrol Process,* **18(10):** 1959-1966.

[4] Kalman, R.E., 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, March 82 (Series D).

[5] Chui, C.K., Chen, G., 1999. *Kalman Filtering with Real Time Applications. third edition*, Springer-Verlag.

[6] RR Karri, A Badwe, et.al. 2013, Application of data assimilation for improving forecast of water levels and residual currents in Singapore regional waters, *Ocean Dynamics* 63 (1), 43-61.

[7] Madsen, H., Canizares, R., 1999. Comparison of extended and ensemble Kalman filters for data assimilation in coastal area modelling. *International Journal for Numerical Methods in Fluids*, **31(6)**: 961-981.

[8]Babovic, V., Fuhrman, D.R., 2002. Data assimilation of local model error forecasts in a deterministic model. *Int. J. Numer. Meth. Fluids*, 39(10): 887-918.

[9]Babovic, V., Sannasiraj, S.A., Chan, E.S., 2005. Error correction of a predictive ocean wave model using local model approximation. Journal of Marine Systems, 53(1-4): 1-17.

[10] Lefohn, A.S., Knudsen, H.P., and Mcevoy, L.R., The Use of Kriging to Estimate Monthly Ozone Exposure Parameters for the Southeastern United-States. *Environmental Pollution,* 1988. **53**(1-4): p. 27-42.

[11] Webster, R. and Oliver, M.A., *Geostatistics for Environmental Scientists 2001*, Chichester, UK: Wiley.

[12] Wang, X and Babovic, V, 2014, Enhancing water level prediction through model residual correction based on Chaos theory and Kriging*, Int. J. Numer. Meth. Fluids* , **75**:42–62

[13] Mancarella, D., Babovic, V., Keijzer, M., Simeone, V., 2008. Data assimilation of forecasted errors in hydrodynamic models using inter-model correlations. *International Journal for Numerical Methods in Fluids*, **56(6)**: 587-605.

[14] Wang, X., Raghuraj, R., Babovic, V., Gerritsen, H., 2010. Artificial Neural Network as a data assimilation tool for error distribution and correction, *9th International Conference on Hydroinformatics*, Tianjin, China.

[15] Kurniawan, A., Ooi, S.K., Hummel, S., Gerritsen, H., 2011. Sensitivity analysis of the tidal representation in Singapore Regional Waters in a data assimilation environment. *Ocean Dynam*, 61(8): 1121-1136.

[16] V Babovic, SA Sannasiraj, ES Chan, 2005, Error correction of a predictive ocean wave model using local model approximation, *Journal of Marine Systems* 53 (1), 1-17.