Review of: "Machine learning for manually-measured water quality prediction in fish farming"

SALIM HEDDAM

Potential competing interests: The author(s) declared that no potential competing interests exist.

In the recently published paper, Zambrano et al. (2021) used two machines learning, i.e., random forests regression (RFR), artificial neural networks (ANN), and the standalone linear regression (LR), for modelling and forecasting several water quality variables in fish farming pond. Two scenarios were analyzed: (*i*) variable estimation and (*ii*) variables forecasting. For the first scenario, an approximation function was established linking one variable to another's, for which, dissolved oxygen concentration (DO), water temperature (T_w), water pH, electrical conductivity (EC), ammonia (NH_3), ammonium (NH_4), nitrites (N_i), and alkalinity (Ak) were modeled separately. For the second scenario (i.e., variables forecasting), each water quality variables was forecasted using the values of the same variable measured at several previous lags times.

In depth review of the present paper reveals several issues that need to be appropriately addressed by the authors.

1. The paper needs to be checked before the final publication. For example, figures 2 and 3 were incorrectly reversed. Indeed, figure 2 should be 3, and vice versa.
2. The introduction lack suffcient information's and discussion to what is already done in relation to the focus of the paper and further complicated by a missing research gap. Appropriate literature review with clearer explanation of principles, ideas, and how machines learning models were applied for modelling and forecasting water quality variables can help in better understanding the objectives of the present investigations.
3. Section data collection is little bit long. In one hand, more detail about the variation and measurement of water quality variables was provided, and in the other hand, the necessary information's for building machine learning models are missing. A table showing the statistical parameters for each variable can help in better highlighting the trend and variation of the water quality variables. In addition, the number of data used for training and validation should be provided, which is actually missing.
4. Models structures are unclear. For the first scenario what were the inputs for each model? For example, according to Table 2, pond T_w is unlikely to have no effect on dissolved oxygen, and it was not considered as an important feature, which seem to be contradictory to what is already reported in the literature where DO and T_w were highly and inversely correlated.

5. For the variables forecasting scenario, without the partial *autocorrelation function (PACF) and the autocorrelation function (ACF),* it would be hard for anyone to know the exact number of lag time were suitable to be included as input variables for the models. For example, in Table 3, very poor forecasting accuracies ($R^2 < 0.80$) was obtained for pond temperature using only one input variable (PO=1). It will be of interest for readers to understand why only one lag time was considered, and if using more than one lag time help in improving the models performances.
6. For more clarity, results using machine learning models should be presented separately for training and validation stage.