17th International Conference on Environmental Science and Technology

Athens, Greece, 1 to 4 September 2021

A data-driven approach to predict phytoplankton blooms using satellite-derived water quality and hydrometeorological drivers

KANDRIS K.[1, *], ROMAS E.[1], TZIMAS A.[1], BRESCIANI M.[2], GIARDINO C.[2], BAUER P.[3], PECHLIVANIDIS I.[4] and DESSENA M.A.[5]

[1] Emvis Consulting Engineers SA, Paparrigopoulou 21, Ag. Paraskevi 153 43, Greece

[2] CNR-IREA, Via Bassini 15, 20133 Milano, Italy

[3] EOMAP GmbH & Co.KG, Schlosshof 4, 82229-Seefeld, Germany

[4] Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

[5] Ente Acque Della Sardegna, Via G Mameli 88, 09123, Cagliari, Italy

*corresponding author

e-mail: kkandris@emvis.gr

Abstract The present work leverages simulated hydrometeorological factors and satellite-derived chlorophyll-a to predict phytoplankton dynamics for Mulargia reservoir (Sardinia, Italy). A Random Forest (RF) model was (a) calibrated to minimize out-of-bag errors of chlorophyll-a predictions for a 5-year-long period (2015-2019), and (b) benchmarked against a naïve predicting alternative for multiple forecasting horizons. Calibration and benchmarking revealed that the RF model can predict the temporal dynamics of phytoplankton growth accurately up to ten days in advance (mean absolute scaled errors ranged from 0.5 to 0.9). Permutation variable importance metrics and the individual conditional expectation plots revealed that moderate temperatures, high soluble phosphorus loads, and low light intensities favor the occurrence of phytoplankton growth in Mulargia. This finding is consistent with the dominance of *Planktothrix* sp. that has been observed in the reservoir. Conclusively, this work lays the foundation for an operational forecast model to help local stakeholders with the present and future reservoir management.

Keywords: Machine learning; forecasting; phytoplankton blooms; remote sensing; hydrometeorological predictions

1. Introduction

Eutrophication is a threat for the ecological state of lakes and reservoirs worldwide. During the last decades, eutrophication has evolved into a significant socioeconomical problem with implications that go beyond its apparent ecological aspects. The management solutions needed to control eutrophication and its impacts are largely supported by scientific results, many of which derive from modelling approaches.

These modeling approaches are generally classified in two classes: empirical (or data-driven) models and mechanistic models (or process-based models) (Vinçon-Leite and Casenave, 2019). Both encompass their own challenges and advantages when used in real-world decision-making processes.

On one hand, mechanistic models can replicate the physical processes involved in eutrophication and, thereby, (a) corroborate or complement monitoring findings, and (b) explore "what if" questions, illuminating specific aspects of eutrophication that could not be otherwise observed. Nonetheless, a meta-analysis, which evaluated the performance of 124 mechanistic models, underscored their difficulties in accurately reproducing phytoplankton dynamics (Shimoda and Arhonditsis, 2016). These difficulties may be, at least partially, attributed to the complex interactions and nonlinear mechanisms that regulate the phytoplankton dynamics, which are modeled via mathematical models based on *a priori* assumptions. Consequently, model-based outcomes become highly uncertain (Gal et al., 2014) and, therefore, questions concerning their application in decision making have been raised (Fornarelli et al., 2013).

On the other hand, data-driven models are simpler in nature, as they require little *a priori* knowledge about the ecosystem processes. Data-driven models have been proven as potent tools for short-term (day to weeks) forecasting (Vinçon-Leite and Casenave, 2019; Cruz et al., 2021), when sufficient data are available. Data availability is a key component for the development of reliable empirical models. In many cases the lack of sufficient data becomes a major obstacle.

This obstacle motivated the present work which leverages simulated hydrometeorological data and satellite-derived water quality observations to develop a data-driven approach – a random forest algorithm – for the short-term forecasting of eutrophication dynamics in lakes and reservoirs.

As already mentioned, data-driven approaches have been used in ecological modelling during the last years and, therefore, the aim of the present effort goes beyond the development of a data-driven model. The aim of this work is three-fold. First, this work aims to test the suitability of relevant data sets that are readily available at a global scale: the development of global hydrometeorological models and the existence of satellite-derived information pose a unique opportunity for data-oriented solutions, even in data-scarce water bodies. Second, this effort aspires to evaluate the limits of forecasting capacities using data-driven models compared to a naïve predicting alternative. Third, the present work aspires to test possible windows into the

model's internals to enhance the interpretability of the typically black-box empirical models. All the above, will be tested in Mulargia reservoir located in South Sardinia, Italy.

2. Methods

2.1 Study area

Mulargia is a large and deep reservoir located in Sardinia (Italy). The reservoir covers an area of 12.5 km^2, has a maximum depth of approximately 99 m and a maximum volume of about 347 hm^3. It serves as a drinking water source for a population of 700,000 inhabitants. Mulargia has a trophic level varying between oligotrophy and mesotrophy (Sulis et al., 2014). Following 2000, the reservoir is clearly dominated by cyanobacteria with the highest occurrence being that of the *Planktothrix agardhii-rubescen*s group (e.g. cyanobacteria presented a peak density in 2002 with a dominance of 99%) (Mariani et al., 2015). Another species of cyanobacteria contributing significantly to the density of the entire class is *Microcystis* sp., with an important bloom in 2001 recording a dominance of 80% (Mariani et al., 2015).

2.2 Datasets

A 5-year-long (2015-2019) record of data has been collected for the training of the random forest (RF) algorithm.

The predictors of the model comprise simulated hydrological and meteorological forcings from continental and global scale models. On one hand, hydrological forcings of the model comprise water inflows, water temperature, and nutrient (total nitrogen and total phosphorus) loadings, as predicted on daily time-step by a re-calibrated version of the E-HYPE model (Pechlivanidis and Crochemore, 2018). On the other hand, meteorological external forcings of the model comprise gridded data of wind, air temperature, precipitation, and solar radiation on an hourly time step from the ERA5-Land reanalysis dataset (Muñoz Sabater, 2019).

The target values of the model, i.e. chlorophyll-a concentrations, have been estimated from imagery acquired from multispectral optical sensors onboard of the polar orbiting satellites Landsat 8 and Sentinel-2. For the period 2015–2019, 195 cloud-free satellite images have been processed to obtain maps of chlorophyll-a concentration. The physical methods for the retrieval of water constituents from Sentinel-2 and Landsat 8 are described elsewhere (Heege and Fischer, 2004; Heege et al., 2014). Ultimately, multispectral chlorophyll-a products were validated for the study area, as described by Bresciani et al. (2019).

For the utilization of the model in a forecasting mode, historical, short-to-medium term forecasts of the hydrometeorological data have been collected for the years 2017-2018 from the same data sources employed for the training of the algorithm.

2.3 Data preparation

Gridded data comprising (a) meteorological variables, and (b) satellite-derived chlorophyll-a concentrations were initially formulated as time-series data with a daily time step corresponding to a specific point of interest within the reservoir, i.e., the water abstraction area of the reservoir. Hydrological data were already available as time series with a daily time step (see also Section 2.2).

An integral step of the prediction strategy and, therefore, of the data preparation phase involved the application of a lag and a sliding window method. To provide an example of how data are formulated in the prediction strategy, for the day-ahead prediction with a ten-day sliding window, (a) a one-day lag was applied to the target values (the time series of chlorophyll-a concentrations), and (b) the predictor matrix contained statistical indices of the hydrometeorological inputs of the last ten days (i.e., the width of the moving window). These statistical indices were: (a) the mean and maximum inflows from the upstream catchment area, (b) the mean water temperature of the inflows from the upstream catchment area, (c) the mean mass of inorganic and organic nitrogen entering the reservoir from the upstream catchment area, (d) the mean mass of soluble and particulate phosphorus entering the reservoir from the upstream catchment area, (e) mean and maximum eastward and northward components of the wind, (f) the mean, maximum and minimum air temperature, (h) the cumulative solar radiation, and (i) the cumulative precipitation.

Due to their sparser temporal resolution, satellite-derived time-series contained missing data. A listwise deletion strategy was adopted to handle missing data. In this regard, cases with missing data were simply omitted and the remaining data were analyzed. Listwise deletion was a reasonable strategy, as it is knowingly producing unbiased estimates and conservative results, since (a) a large enough sample of data remained, and (b) the assumption of random missingness was satisfied.

2.4 Ensemble learning algorithm

The RF algorithm is an ensemble learning approach that relies on the bootstrap aggregation (bagging) of regression trees with some additional degree of randomization. Bagging of regression trees alleviates their instability, while randomization reduces the correlation among them and, consequently, reduces the variance of the predictions (i.e., the average of the trees). Randomization is conducted by randomly selecting a subset of predictor variables as candidates for splitting. Prediction is performed by averaging the predictions of each tree. Tyralis et al. (2019) offer a comprehensive review on the formulation of RF in water resources applications.

Four hyperparameters of the RF algorithm, i.e., the minimum leaf size of each tree, the number of splits, the number of trees, and the number of predictor variables to sample, were fine-tuned using a Bayesian optimization method to minimize out-of-bag errors of chlorophyll-a predictions in Mulargia reservoir. Out-of-bag errors correspond to the samples (about 1/3 of the training set)

that remain after bootstrapping the training set. In addition to the RF parameters, the optimum width of the sliding window for the prediction strategy was investigated. Windows varying from one to ten days were tested.

As mentioned in the Introduction, this work searches for windows into the model's internals to enhance its otherwise limited interpretability. These windows are two-fold: (a) a permutation variable importance metric (VIM), and (b) a model-agnostic visualization tool, i.e., the individual conditional expectation (ICE) plot. On one hand, a permutation VIM measures the mean decrease in accuracy in the out-of-bag sample by randomly permuting the predictor variable of interest. If a predictor is influential in prediction, then permuting its values should drop model accuracy and high VIM values should be anticipated. On the contrary, if a predictor is not influential, then permuting its values should have little to no effect on the model error and VIMs will be randomly distributed around zero. On the other hand, ICE plots expand and refine the classical partial dependence plots (PDP) by graphing the functional relationship between the predicted response and the feature for individual observations. ICE plots highlight the variation in the fitted values across the range of a covariate, suggesting where and to what extent heterogeneities might exist.

2.5 Benchmarking

To evaluate the predictive power of the RF algorithm, predictions from the RF algorithm were benchmarked against a naïve method, i.e., the last observation method. The metric used for the comparison was the mean absolute scaled error (MASE), which is the mean absolute error of the forecast values, divided by the mean absolute error of the in-sample one-step naive forecast (Hydnman, 2006). The MASE was estimated for increasing forecasting horizons (i.e., moving from day-ahead to ten-day-ahead predictions) for the years 2017-2018. For these two years, the model employed expired real-time forecasted hydrometeorological data rather than the re-analysis and historical simulated datasets that it has been trained on.

3. Results and discussion

3.1. Algorithm development and performance

The Bayesian optimization algorithm reached a solution within 67 iterations with an out-of-bag error equal to 4.1 μg chlorophyll-a/l. A 10-day sliding window for both forcing data types (hydrological and meteorological) was optimal; in other words, the hydrometeorological data for the last 10 days are required for predicting day-ahead chlorophyll-a values. Regarding the formulation of the RF algorithm, the optimal solution of the optimization problem corresponded to a minimum leaf size of 5, 13 variables to sample, 14 splits and 329 regression trees.

Figure 1 demonstrates the output of the parameter fine-tuning effort and compares the day-ahead predictions of the model with the satellite-derived observations of chlorophyll-a for the training part of the dataset. The model managed to capture accurately the temporal dynamics of phytoplankton growth, achieving a mean absolute error of 2.8 μg/l. The model predicted accurately the timing of chlorophyll-a spikes, but slightly underpredicted their intensity.

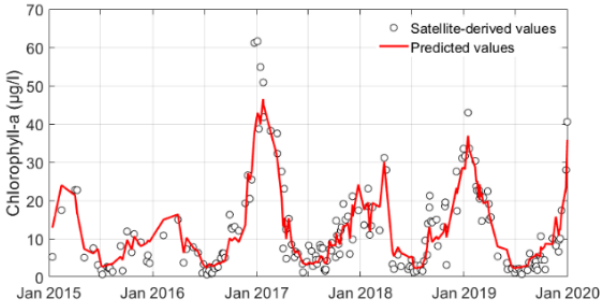

Figure 1. Observed versus simulated values for the RF model for the training data (2015-2019).

When tested in a forecasted rationale and for larger forecasting horizons, the accuracy of the model gradually deteriorated, but remained superior to its naïve alternative until a ten-day-ahead forecasting horizon, as shown in Figure 2; MASE remained lower than one for all forecasting horizons.

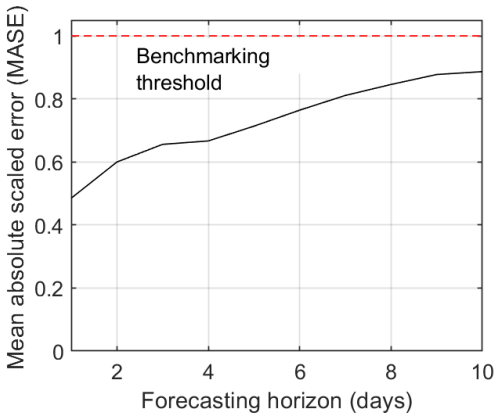

Figure 2. MASE for the RF model as a function of the forecasting horizon for the years 2017-2018.

3.2. Predictor importance assessment

Results based on the permutation VIM adopted herein, revealed that air temperature, cumulative radiation, and total soluble phosphorus mass entering the reservoir were the most influential predictors, followed by the temperature of inflows (see Figure 3). On the contrary, predictors related with wind, the total precipitation or the nitrogen-related loads could be excluded from the predictor list, as they exhibited VIMs close to zero. To further get an inkling on the relevance of the most influential parameters, their individual expectation plots indicated that moderate temperatures (Figure 4a and b), high soluble phosphorus loads (Figure 4c) and low light intensity (Figure 4d) are favoring chlorophyll-a production in Mulargia. This finding is consistent with the dominance of *Planktothrix* sp. that has been observed in the reservoir.

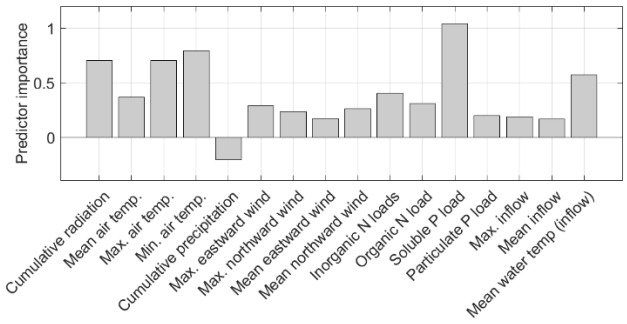

Figure 3. Predictor importance estimates for the RF algorithm.

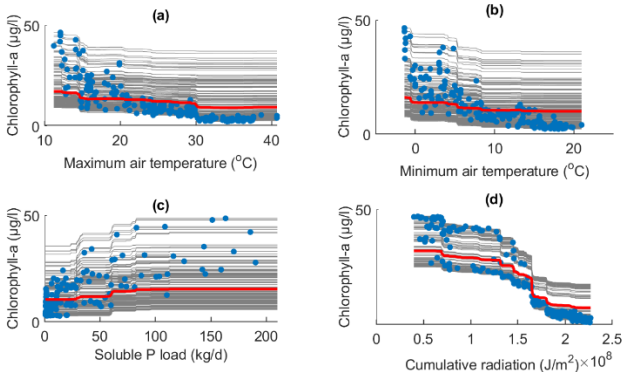

Figure 4. ICE plots for (a) maximum air temperature, (b) minimum air temperature, (c) mean soluble phosphorus load, (d) cumulative radiation. The red line shows the PDP; the gray lines and blue points are derived from the ICE analyses.

4. Conclusions

This work developed and evaluated the predictive abilities of an RF algorithm that employs simulated hydrometeorological drivers to predict previously validated satellite-derived chlorophyll-a concentrations. Results indicate that the RF algorithm adopted herein could capture the dynamics of phytoplankton growth adequately. When benchmarked with a naïve forecasting alternative, the RF provided accurate forecasts for up to ten days in advance.

Besides model evaluation, two complementary ways to gain insight in the model's internals were used to augment the understanding of what drives phytoplankton growth in the reservoir. Using the VIM and ICE plots, moderate temperatures, high soluble phosphorus concentrations, and low light intensity were found to favor phytoplankton growth.

Perhaps more importantly, this work lays the foundation for an operational forecast model to help local stakeholders with the present and future reservoir management. The type of data allows to use the same approach to other inland water bodies regardless of the availability of regional in-situ data. In what follows this part of the work, the application of this methodology in diverse water bodies will be performed to confirm its generalization potential.

References

Bresciani, M., Giardino, C., Stroppiana, D., Dessena, M. A., Buscarinu, P., Cabras, L., ... & Tzimas, A. (2019). Monitoring water quality in two dammed reservoirs from multispectral satellite data. European Journal of Remote Sensing, 52(sup4), 113-122.

Cruz, R. C., Reis Costa, P., Vinga, S., Krippahl, L., & Lopes, M. B. (2021). A Review of Recent Machine Learning Advances for Forecasting Harmful Algal Blooms and Shellfish Contamination. Journal of Marine Science and Engineering, 9(3), 283.

Fornarelli, R., Galelli, S., Castelletti, A., Antenucci, J. P., & Marti, C. L. (2013). An empirical modeling approach to predict and understand phytoplankton dynamics in a reservoir affected by interbasin water transfers. Water Resources Research, 49(6), 3626-3641.

Gal, G., Makler-Pick, V., & Shachar, N. (2014). Dealing with uncertainty in ecosystem model scenarios: application of the single-model ensemble approach. Environmental modelling & software, 61, 360-370.

Heege, T., & Fischer, J. (2004). Mapping of water constituents in Lake Constance using multispectral airborne scanner data and a physically based processing scheme. Canadian Journal of Remote Sensing, 30(1), 77-86.

Heege, T., Kiselev, V., Wettle, M., & Hung, N. N. (2014). Operational multi-sensor monitoring of turbidity for the entire Mekong Delta. International Journal of Remote Sensing, 35(8), 2910-2926.

Hyndman, R. J. (2006). Another look at forecast-accuracy metrics for intermittent demand. Foresight: The International Journal of Applied Forecasting, 4(4), 43-46.

Mariani, M. A., Padedda, B. M., Kaštovský, J., Buscarinu, P., Sechi, N., Virdis, T., & Lugliè, A. (2015). Effects of trophic status on microcystin production and the dominance of cyanobacteria in the phytoplankton assemblage of Mediterranean reservoirs. Scientific reports, 5(1), 1-16.

Muñoz Sabater, J., (2019): ERA5-Land hourly data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed on 10-Nov-2020), 10.24381/cds.e2161bac

Pechlivanidis, R. & Crochemore L., (2018) Validation report for hydrological model service. Deliverable 6.3 of the Space-O project funded under the European Union's Horizon 2020 research and innovation programme GA No: 730005. Available at: https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5bd50fb15&appId=PPGMS

Shimoda, Y., & Arhonditsis, G. B. (2016). Phytoplankton functional type modelling: Running before we can walk? A critical evaluation of the current state of knowledge. Ecological modelling, 320, 29-43.

Sulis, A., Buscarinu, P., Soru, O., & Sechi, G. M. (2014). Trophic state and toxic cyanobacteria density in optimization modeling of multi-reservoir water resource systems. Toxins, 6(4), 1366-1384.

Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water, 11(5), 910.

Vinçon-Leite, B., & Casenave, C. (2019). Modelling eutrophication in lake ecosystems: a review. Science of the total environment, 651, 2985-3001.