A spatiotemporal synthesis of high-resolution salinity data with aquaculture applications.

Dong Liang
University of Maryland Center for Environmental Science Chesapeake Biological Laboratory
Solomons MD 20688
dliang@umces.edu

Jeremy M. Testa
University of Maryland Center for Environmental Science Chesapeake Biological Laboratory
Solomons MD 20688
jtesta@umces.edu

Cassie Gurbisz
St. Mary's College of Maryland
Marine Science Department
St. Mary's City, MD 20686
cbgurbisz@smcm.edu

Lora A. Harris
University of Maryland Center for Environmental Science Chesapeake Biological Laboratory
Solomons MD 20688
harris@umces.edu

***Abstract*— Technological advancement and the desire to better monitor shallow habitats in the Chesapeake Bay, Maryland, United States led to the initiation of several high-resolution monitoring programs such as ConMon (short for "Continuous Monitoring") measuring oxygen, salinity, and chlorophyll-a at a 15-minute frequency. These monitoring efforts have yielded an enormous volume of data and insight into the condition of the tidal water of the Bay. But this information is underutilized in documenting the fine-scale variability of water quality, which is critical in identifying the link between water quality and ecological responses, partly due to the challenges in integrating monitoring data collected at different frequencies and locations. In a project to understand the environmental suitability of aquaculture sites and the future potential overlap between aquaculture and submerged aquatic vegetation, we developed a spatiotemporal synthesis of ConMon data with data from long-term, fixed-station seasonal monitoring. Here, we present our generalized additive model-based approach to predict salinity at high frequency (15 minutes) and fine spatial resolution (~100 meters) in the Maryland portion of the Bay, its major tributaries, and the shallow tidal creeks that exchange with the tributaries. Predictive performance was validated to be 1 PSU (practical salinity unit) in root mean square error using de novo monitoring. The resulting data provide insights into the environmental suitability of aquaculture, specifically the sensitivity of the Easter oyster (*Crassostrea virginica*) to low salinity stress. The spatiotemporal synthesis approach has potential applications for integrated monitoring and potential linkage with high-resolution water quality models for shallow habitats.**

Keywords— big data, data fusion, hydroinformatics, salinity estimates.

I. INTRODUCTION

The convergence of data science, computing (e.g. parallel computing and distributed storage systems) and hydrology [1] have enabled analyses using big data in hydroinformatics. The combination of open data policies and cloud computing now allows for the merging of high-quality data from agencies with advanced cyberinfrastructure and technical capacity for data management and analysis [2]. While the open-data trend enables data sharing and analysis, other issues such as data format, data resolution, and metadata development require collaboration and innovative solutions[3].

Synthesis of location-specific environmental data with differing temporal and spatial resolution has been a problem in multiple domains: remote sensing [4], atmospheric science [5], and hydroinformatics[6]. Common approaches include geostatistical approaches such as Kriging with external drift and co-Kriging [7,8]. These approaches estimate the correlation between different models of data, and spatial cross-covariance of measured quality to enhance prediction. But kriging tends to be limited by sparse data. A different approach builds flexible predictive models for higher-quality sparse data with lower-quality complete data such as geographically weighted regression (GWR) or artificial neural networks [9]. This approach can easily incorporate ancillary data and the time domain to address the sparse data issues. But model fitting and optimization can be time-consuming. Furthermore, all of these approaches can be computationally challenging for big data.

This case study unites disparate high spatial-resolution and high temporal-resolution salinity datasets from the Chesapeake Bay. Continuous Monitoring (ConMon) is a spatially sparse, but high temporal resolution monitoring program for shallow habitats in the Chesapeake Bay. Such data is valuable to document the fine-scale variability of water quality at specific locations and understand the environmental suitability of aquaculture sites and future potential overlap between aquaculture and submerged aquatic vegetation [10]. We developed a statistical framework to fuse the ConMon station records with complementary data sources to build a salinity record with high spatial resolution and high temporal frequency. These salinity records are useful to understand the vulnerability of locations to low-salinity conditions, which impact the site suitability for growing oysters.

II. MATERIALS AND METHODS

A. Study Area

The Chesapeake Bay system (denoted Bay hereafter) is a large estuary located on the Mid-Atlantic coast of the USA. The Bay is shallow and connected with over 60 tributaries, with a mean freshwater discharge of 2219 m^3s^{-1}, and a freshwater fill time of about one year. The ratio of drainage basin area to estuarine surface area for the entire Bay is 14:1, which indicates the potential for large terrestrial influence. Freshwater inputs to the Bay are a primary driver of many key physical and biological

processes, including vertical stratification in the summer [11], summer hypoxia [12], and strong gradients in salinity [13].

B. Data

The long-term Chesapeake Bay monitoring program is a comprehensive water quality and habitat monitoring program. The program collects data at 133 stations in the Bay; bimonthly in warmer months (May through September) and monthly throughout the year, with vertical profiles collected at 1-meter resolution. The program has assessed water quality since 1984 by measuring nutrients and habitat conditions such as salinity, dissolved oxygen, and water clarity. The shallow water monitoring program deploys continuous monitoring (ConMon) at over 80 locations in the Bay, some of which are maintained as sentinel locations and others that are rotated between locations every 3 years. Widely available multi-sensor probes (YSI, Xylem) are deployed to collect data continuously every 15 minutes for habitat conditions such as surface salinity, surface temperature, and dissolved oxygen (Fig. 1).

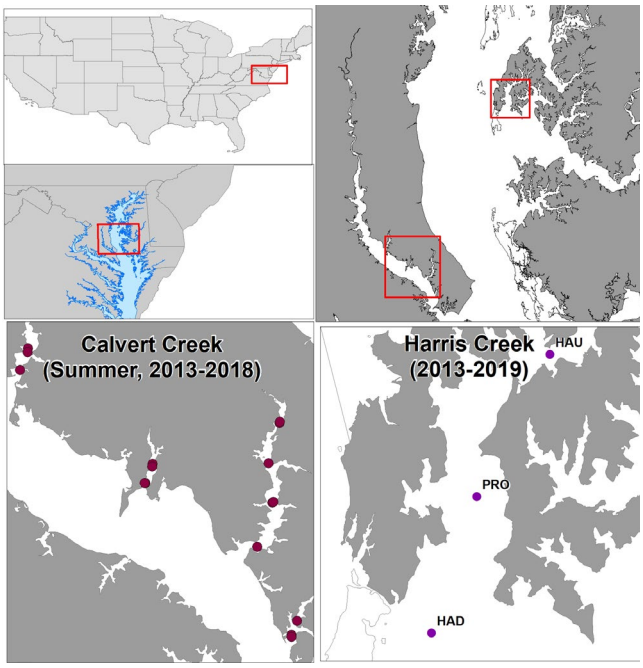

Figure 1 Calvert Creek and Harris Creek sampling stations and duration in Maryland, United States. Calvert Creek sampling follows fortnightly design while Harris Creek sampling utilized the continuous monitoring schemes.

C. Data Processing

To prepare the data for processing, long-term stations were divided into 48 Bay segments, and grouped these into 13 batches. Each batch consists of seven years when contemporary aquaculture data are available. We then linked surface grab samples from long-term stations to the nearest ConMon station and matched observations within 1 hour of the grab sonde reading. Average hourly sonde measurements were computed for each Julian day for each ConMon station. The matched ConMon data represented the expected high-frequency seasonal variation of salinity for each long-term station. The matching process was conducted separately for each segment in the Maryland portion of the Bay, using separate polygons to delineate the search areas and to implement neighbor searches using water distance instead of geographic distance. Each batch was processed in parallel using separate CPUs.

D. Statistical Framework

Within each segment and year, we formulate several Geographically Weighted Regressions (GWR) to predict salinity, with the following formulation.

$$z_i \sim \beta_0 + \mathrm{te}(x_i, y_i) + \mathrm{s}(j_i, bs = \mathrm{cc}) + \mathrm{te}(x_i, y_i, \mathrm{by} = u_i) + \epsilon_i$$

where z is measured salinity at a long-term station, x, and y are the longitude and latitude of the measurement, j is the Julian day (i.e. day of year), and u is the average hourly salinity from the nearest ConMon station on the same Julian day. Parameters include β_0 the intercept, s denotes a cubic regression spline, bs=cc denotes cyclic cubic regression spline, te denotes the bivariate thin-plate spline built through tensor products for the slopes in average salinity defined via "by=u", and ϵ denotes the residuals following a normal distribution. The model was implemented in the *mgcv* package in R [14], and run on a Beowulf cluster with 96 cores and 256 Gb RAM, and 1 Tb storage. Hourly prediction generated intermediate files approaching 10 Gb for big segments. For these segments, the predictive analysis took around 55.6 hours.

III. Results

Three GWR models were evaluated using ten-fold cross-validation. The long-term monitoring data were divided randomly into 10 batches. The models were trained with 9 batches and tested on the remaining batch iteratively (Fig. 2). All long-term data were used in the cross-validation (n=1,055,385). We tested a GWR model with spatially varying coefficients, and two dynamic GWR models with daily, or hourly-specific varying coefficients. Model skill was visualized using smoothed scatter plots. Cross-validation suggests a

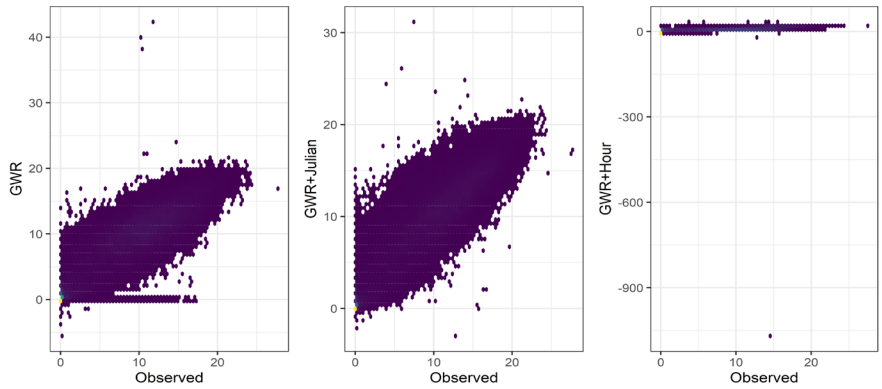

Figure 2: Ten-fold cross validation results based on long term monitoring data (n= 1,055,385) for three Geographically Weighted Regression predictions for salinity. Left: spatially varying coefficients, Middle: spatially and daily (Julian) varying coefficients. Right: spatially and hourly varying coefficients.

balance between model complexity and predictive performance. The spatial-only GWR model predicted less well, while the hourly GWR was over-parameterized and generated spurious results. The dynamic GWR with medium complexity generated the best overall predictive performance and was therefore selected for further evaluation.

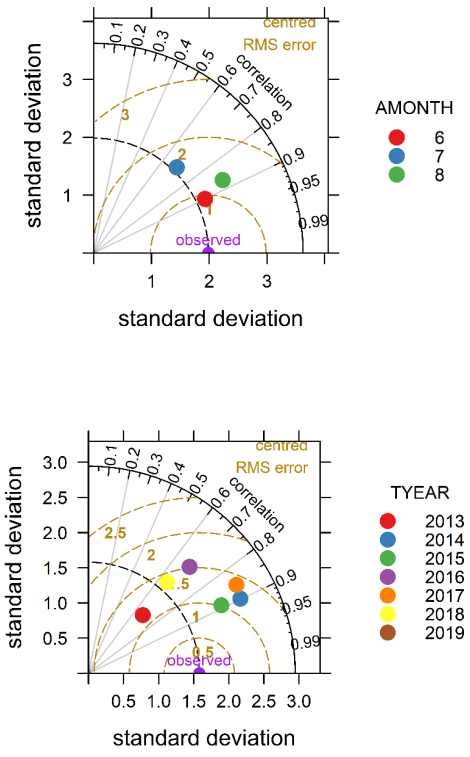

Figure 3: Evaluation of model prediction based on theTop: Calvert Creek Monitoring study, Bottom: High frequency time series of sampled and fused salinity from Harris Creek between 2013-2019 via Taylor diagrams for hourly salinity prediction.

The predictive performance of the proposed models was evaluated with a portion of monitoring data collected as part of a county-supported monitoring program in the tidal tributaries of the Patuxent River estuary (itself a tributary of Chesapeake Bay). These data, which were collected as part of a program called "Calvert Creeks", were not used in training the GWR. The sampling design of Calvert Creek monitoring is similar to long-term monitoring, except that data are only collected during the warm season (May to September) and only surface and bottom measurements are made. The hourly GWR prediction was aligned with the nearest monitoring data over time and space (Fig. 3). Predictive performance was validated to be 1 PSU (practical salinity unit) in root mean square error using Calvert Creek monitoring. The prediction was slightly less accurate in July than in June and August. An alternative ConMon dataset collected in Harris Creek (Fig. 1) was also used to validate the models. Predictive performance was

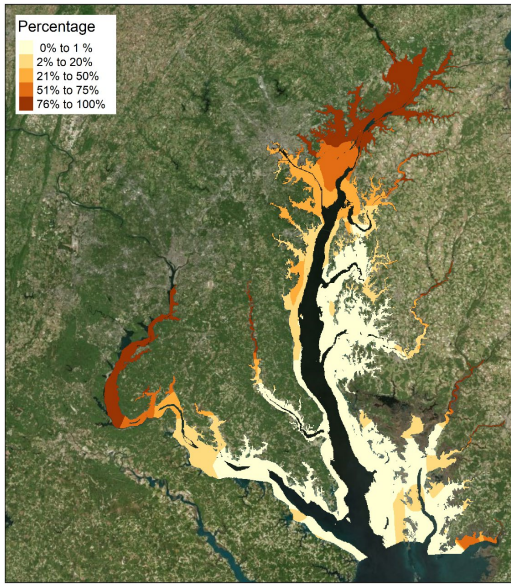

Figure 4: Average of seven monthly proportions (April through October) of hours with predicted low salinity (<5) at 100m resolution. Salinity was predicted using a geographically weighted regression of long term monitoring data and continuously monitoring data. Each monthly proportion is a cell and month specific proportion of hours between 2012-2019 with low salinity.

approximately 1 PSU in root mean square error (RMSE) using high-frequency time series of the monitoring.

A monthly median of hourly salinity, as well as extremes (lower and upper 5 percentiles), were predicted using the best-performing model (Fig. 4). The hourly salinity was predicted at the centroid of each raster cell for each year between 2012 and 2019. The hourly predictions from each period were temporally aggregated over years into long-term summaries. Missing values due to insufficient nearest ConMon data were interpolated using inverse distance weighting.

IV. Discussion and Conclusion

Shallow habitats exhibit fine-scale variability of water quality, which is a challenge for agencies interested in monitoring water quality criteria and for aquaculturists interested in choosing optimal locations for their farms. To understand the environmental suitability of aquaculture sites and the future potential overlap between aquaculture and submerged aquatic vegetation, a Geographically Weighted Regression was used to develop a high-spatial resolution salinity dataset by combining ConMon and standard monitoring data. This statistical method generates hourly predictions of salinity within ~100 m grid cells. Validation studies suggest reasonable accuracy of the predictions in terms of RMSE in regions of de novo sampling. Data fusion, as demonstrated in this study, can significantly enhance our ability to characterize the spatiotemporal variation of shallow habitats, which is critical in identifying the link between water quality and ecological response.

Our analyses can be improved in the following aspects. First data fusion is challenging in sparsely sampled regions. The performance of other methods (e.g. Geostatistical fusion) remains to be evaluated. In particular, Bayesian principles can be applied to build prior from low-quality complete data and correct it with higher-quality sparse observations [15,16]. Multivariate approaches such as principle component analysis, and wavelet analyses, may also be applied to this problem [17].

Water quality modeling can benefit from cyber-infrastructure development to enable a scalable and reproducible workflow [18]. In this data synthesis effort, gigabytes of time series data were manually processed and integrated. The synthesis process could be automated to make datasets for similar shallow habitat parameters, such as oxygen and chlorophyll-a, available in high resolution to the broader research community. Analysis conducted at high resolution (e.g. 100-m resolution over 15-minute frequency for Chesapeake Bay between 2013-2019) led to a computing bottleneck during the predictive modeling and optimization stage. The current workflow took more than 24 hours, which may not be ideal for some use cases that require a faster turn-around [2]. High-performance computing platforms can be utilized to potentially alleviate the computing bottleneck and produce similar results but within a more reasonable turnaround time. Figures and Tables

ACKNOWLEDGMENT

We thank the Maryland Department of Natural Resources for collecting the data. We also thank Calvert County, Maryland for supporting the monitoring program. This research was funded by the National Oceanic and Atmospheric Administration through the Maryland Sea Grant (MSDG Omnibus Award # 07528187). This is UMCES Contribution No. 6332.

REFERENCES

[1] Chen, Y. and D. Han, Big data and hydroinformatics. Journal of Hydroinformatics, 2016. 18(4): p. 599-614.

[2] Liu, Y.Y., et al., A CyberGIS integration and computation framework for high‐resolution continental‐scale flood inundation mapping. JAWRA Journal of the American Water Resources Association, 2018. 54(4): p. 770-784.

[3] Xu, H., et al., An overview of visualization and visual analytics applications in water resources management. Environmental Modelling & Software, 2022: p. 105396.

[4] Zhu, X., et al., An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. Remote Sensing of Environment, 2010. 114(11): p. 2610-2623.

[5] Liang, D. and N. Kumar, Time-space Kriging to address the spatiotemporal misalignment in the large datasets. Atmospheric Environment, 2013. 72: p. 60-69.

[6] Hu, Q., et al., Rainfall spatial estimations: A review from spatial interpolation to multi-source data merging. Water, 2019. 11(3): p. 579.

[7] Velasco-Forero, C.A., et al., A non-parametric automatic blending methodology to estimate rainfall fields from rain gauge and radar data. Advances in Water Resources, 2009. 32(7): p. 986-1002.

[8] Cecinati, F., et al., Representing radar rainfall uncertainty with ensembles based on a time-variant geostatistical error modelling approach. Journal of Hydrology, 2017. 548: p. 391-405.

[9] Wehbe, Y., M. Temimi, and R.F. Adler, Enhancing precipitation estimates through the fusion of weather radar, satellite retrievals, and surface parameters. Remote Sensing, 2020. 12(8): p. 1342.

[10] Dumbauld, B.R. and L.M. McCoy, Effect of oyster aquaculture on seagrass Zostera marina at the estuarine landscape scale in Willapa Bay, Washington (USA). Aquaculture Environment Interactions, 2015. 7(1): p. 29-47.

[11] Murphy, R.R., W.M. Kemp, and W.P. Ball, Long-term trends in Chesapeake Bay seasonal hypoxia, stratification, and nutrient loading. Estuaries and Coasts, 2011. 34: p. 1293-1309.

[12] Scavia, D., et al., Advancing estuarine ecological forecasts: seasonal hypoxia in Chesapeake Bay. Ecological Applications, 2021. 31(6): p. e02384.

[13] Chatwin, P., Some remarks on the maintenance of the salinity distribution in estuaries. Estuarine and Coastal Marine Science, 1976. 4(5): p. 555-566.

[14] Wood, S.N., Generalized additive models: an introduction with R. 2017: CRC press.

[15] Ma, Y., et al., Performance of optimally merged multisatellite precipitation products using the dynamic Bayesian model averaging scheme over the Tibetan Plateau. Journal of Geophysical Research: Atmospheres, 2018. 123(2): p. 814-834.

[16] Verdin, A., et al., AB ayesian kriging approach for blending satellite and ground precipitation observations. Water Resources Research, 2015. 51(2): p. 908-921.

[17] Kalinga, O.A. and T.Y. Gan, Merging WSR-88D stage III radar rainfall data with rain gauge measurements using wavelet analysis. International journal of remote sensing, 2012. 33(4): p. 1078-1105.

[18] Choi, Y.-D., et al., Toward open and reproducible environmental modeling by integrating online data repositories, computational environments, and model Application Programming Interfaces. Environmental Modelling & Software, 2021. 135: p. 104888.