Geographically aware estimates of remotely sensed water properties for Chesapeake Bay

Jemma Stachelek[1], Sofia Avendaño[1,2], Jon Schwenk[1]

[1]Los Alamos National Laboratory, Division of Earth and Environmental Sciences, Los Alamos, NM 87545, USA

[2]Nicholas School of the Environment, Duke University, Durham, NC 27708, USA

Abstract

Remotely sensed water properties are important for a variety of applications, including validation of Earth Systems models (ESMs), habitat suitability models, and sea level rise projections. For the validation of next-generation, high or multi-resolution (30-60 km) ESMs in particular, the usefulness of operational forecasting products and directly-observing satellite-based sensors for validation is limited due to their temporal availability and spatial resolution of < 1 year (in some cases) and $> 30\ km^2$ respectively. To address this validation data gap, we developed a data-driven model to produce high-resolution ($< 1\ km^2$) estimates of temperature, salinity, and turbidity over decadal time scales as required by next-generation ESMs. Our model fits daily MODIS Aqua reflectance data to surface observations (< 1 m depth) from 2000-2021 in Chesapeake Bay, USA. The resulting models have similar error statistics as prior efforts of this type for salinity (RMSE: 2.3) and temperature (RMSE: 1.8 C). However, unlike prior efforts our model is designed as a pipeline meaning that it has the advantage of producing predictions of water properties in future time periods as additional MODIS data becomes available. We also include novel "geographically-aware" predictive features insofar as they capture geographic variation in the influence of flow and surface water exchange in upstream coastal watersheds.

Keywords: reflectance, remote sensing, salinity, temperature, turbidity, Chesapeake

1 Introduction

In partially mixed estuaries such as the Chesapeake Bay (hereafter, "the Bay"), productivity, organic matter production, and phytoplankton community composition are largely determined by stratification [1]. Because stratification is controlled by salinity and temperature, extensive measurement networks exist to monitor these water properties (Chesapeake Bay Program Water Quality Database, `https://www.chesapeakebay.net/what/downloads/cbp_water_quality_database_1984_present`). These data are critical for establishing point-based trends and relationships between river flow or sea level rise and salinity or hypoxia [2]. Of particular importance is monitoring for increased nutrient loads as result of expected increases in precipitation, tributary discharge, and river flows as a result of climate change [3, 4].

Unfortunately, direct observations alone are not adequate for many applications that require spatially extensive data such as Sea Level Rise (SLR) forecasting and habitat suitability assessments [5]. For example, habitat suitability assessments are typically done on a gridded basis for the entirety of a particular estuary rather than at selected point locations. Applications such as these often turn to gridded alternatives to direct observations such as output from directly-observing satellite-based sensors [6] or operational now/fore-casting products [7]. Unfortunately, these too are of limited use for applications such as validation of next-generation Earth Systems Models (ESMs, [8, 9]), where the temporal availability of operational forecasting products is too short (< 1 year) and the spatial resolution of directly-observing satellite-based sensors is too coarse ($> 30\ km^2$). There is a strong need in particular for satellite-based products to validate recently developed biogeochemical modules of the E3SM Earth Systems model which aims to represent thermal, sediment, salinity, nutrient exchanges between rivers and the land on a global basis [9].

To address this validation data gap, we developed a data-driven model to produce high-resolution (¡ $1\ km^2$) estimates of temperature, salinity, and turbidity at daily time-steps (over decadal time scales) as

required by next-generation ESMs. The primary input to our model is remote sensing reflectance data. In particular, we use Moderate Resolution Imaging Spectroradiometer (MODIS) Aqua multiband data, which has been shown to be predictive of salinity in coastal Louisiana [10, 11]. We supplemented MODIS data with gridded estimates of distance-weighted tributary inflows which we used to encourage physical realism (i.e. geographic "awareness") especially in areas of the Bay with little to no observations. Our model was trained against not only buoy observations taken in the mainstem of the Bay but also observations taken in the mid to lower reaches of Bay tributaries. These more spatially extensive observations likely increase the accuracy and reliability of model predictions in the upper tributaries.

2 Methods

2.1 Study site

Chesapeake Bay is located in the Mid-Atlantic region of the United States (at approximately 37.8°N 76.1°W, see Figure S1) and serves as the outlet of 9 major tributary systems. Of these, the Susquehanna River (at the Northern extent of the Bay) provides the majority of the freshwater supply to the Bay (~48%, [1]). The Bay is the largest estuary in the United States and has an area of approximately 11,000 km^2. As such, it is provides a wealth of natural economic benefits in the form of fisheries harvests (particularly oysters), tourism, and nature-based recreation [5]. Despite its large size, the Bay is fairly shallow (< 30m, [12]) and the hydrodynamics of the Bay are such that it is considered stratified and partially mixed [1]. For the present study, this means that surface water properties are likely to differ substantially from water properties at deeper depths. It also means that the Bay receives enough freshwater inputs to produce a distinct upper-estuary to lower-estuary gradient in water properties.

2.2 Model overview

Prior data-driven models (e.g. [13, 10]) of Bay water properties used Generalized Linear Models (GLM), Neural Networks (NN), Generalized Additive Models (GAMs), and Random Forests (RF). We chose to use the Random Forest algorithm in part because [13] found that it produced nearly equivalent results to more complex methods (particularly NN) and also because of its ease of implementation and interpretation. Random Forests have a particular advantage because they are robust against interactions among variables and allow for nonlinearities [14]. We fit RF models using the scikit-learn Python package [15] where each model was initialized with 250 trees, had a maximum depth of 20 tree "layers", and was subjected to recursive feature elimination to produce the most parsimonious model (Figure 1). We chose initial forest size and maximum depth values to optimize model performance and model training time respectively. We subsequently "tuned" each model using k-fold cross validation to optimize the size, depth, and splitting criteria of each tree. We evaluated the "importance" of particular variables as the average impurity decrease of tree splits based on that variable [15].

2.3 Data description

We compiled over two decades of observational data (2000-2021) on salinity (n=6k), temperature (n=10k, in units of °C), and turbidity (n=5k, in units of NTU, See Table S1). Although this data came from a wide variety of sources (n=20, Table S1), almost 80% of the total unique observations came from three programs NDBC/CBIBS (https://buoybay.noaa.gov/), Maryland Continuous Monitoring Program (MDNR), and the Virginia Estuarine and Coastal Observing System (http://vecos.vims.edu). The temporal range of the observations (2000-2021) includes the full range of freshwater input magnitudes to the Bay spanning both times of minimum and maximum freshwater input (https://www.usgs.gov/media/images/estimated-annual-mean-streamflow-entering-chesapeake-bay).

Because Chesapeake Bay has a mean optical depth of approximately 0.89m [16], we filtered all collected observations to those occurring in the top 1m of the water column. We further limited observations to those that had a timestamp during the MODIS overpass window time-of-day (approximately 1700 - 1900 hours GMT). Prior to model fitting, we removed the annual cycle from temperature observations using least squares optimization routines [17]. This annual cycle was added back to model estimates for evaluation purposes

2

70 and in the production of prediction surfaces. As a final observational data processing step, we implemented spatial and temporal averaging of observational data to a daily mean within each MODIS pixel.

72 We supplemented observational data with a grid of "freshwater influence" (FWI) estimates where each grid cell represents the average of annual-flow weighted distance of that cell from the 5 major tributaries

74 (Susquehanna, James, Pautexent, Choptank, and Potomac). Annual flow magnitudes were calculated from National Water Information System instantaneous data (sites: 01488110, 01491000, 01576000, 01654000,

76 01594440, 02037500, [18]). Rather than straight-line distances we used "in-water" distances calculated using the scikit-image Python package [19, 20]. Our in-water distance calculations involved creating a cost–surface

78 whereby land pixels had a very high cost relative to water pixels. Then the distance from any one cell to a tributary source was the accumulated travel "cost" as given by the scikit-image `route_through_array`

80 function. We combined FWI estimates with the pixel-wise daily mean observations and regressed them in RF models against data from the 1000m resolution MODIS Aqua product, which provides 9 bands of values

82 for each pixel spanning from the following wavelength ranges (405-420, 438-448, 483-493, 526-536, 546-556, 662-672, 673-683, 743-753, and 862-877nm) [21]. MODIS data was pulled from Google Earth Engine (Asset:

84 `MODIS/006/MYDOCGA`) using the `earthengine-api` Python package.

All code for data processing, model fitting, and model evaluation for the pipeline shown in Figure 1 is

86 available in [22]. All data used in the study are available in [23].

2.4   Model validation

88 We validated RF model predictions both quantitatively against a hold-out test set of observational data (one third of the total data) and qualitatively against data from the Chesapeake Bay Operational Forecast

90 System (CBOFS, [7]). CBOFS provides data on Bay-wide water level, temperature, and salinity every 6 hours. The model itself uses National Oceanic and Atmospheric Administration (NOAA) water level, salinity

92 observations, and US Geological Survey (USGS) river discharge data to calibrate an implementation of the Rutgers University Regional Ocean Modeling System (ROMS, [24]).

3   Results

Our tuned RF models for salinity, temperature, and turbidity had a root mean square error (RMSE) of 2.34

96 (R^2=0.91), 1.76 (R^2=0.96), and 1.82 (R^2=0.11) respectively (Table 1, Figure 2). Out-of-sample validation performance was good for salinity and temperature models but poor for the turbidity model (Figure 2, Figure

98 S2). Despite excellent overall performance for the temperature and salinity models, there was some indication of poor performance at the extremes. The models were unable to recover very low temperature values ($<$

100 0 C) or very high temperature and salinity values ($>$ 30 C, $>$ 23 salinity, Figure 2). One explanation for these areas of poor performance is that such extremes were rare in the observational dataset (Table S2).

102 The salinity model however did faithfully capture areas of low salinity (Figure 2, S4). These low salinity values can be regarded as low uncertainty (high confidence) estimates given the extent to which the upper

104 tributaries are well represented in the observational dataset (Figure S3).

Table 1: Random Forest model fit and predictive accuracy metrics. (RMSE = root mean square error, R^2 = coefficient of determination)

metric	salinity	temperature	turbidity
RMSE	2.34	1.76	1.82
R^2	0.91	0.96	0.11

The date of observation was the most important feature identified in RF training for the temperature

106 and turbidity models while location of observation was most important in the salinity model (Figure 3). Across all models, the most important MODIS band was band 8, which is the shortest wavelength band of

108 the MODIS Aqua product at the extreme of the visible light range. The "size" (i.e. n_estimators) of the final RF models were of a similar magnitude for each variable but the tree was much "deeper" (i.e. max_depth)

110 for temperature compared to the other variables (Table S3).

Running the models in prediction mode produced surfaces with a realistic spatial trend whereby salinity was highest near the Bay mouth and decreased with distance from each tributary mouth (Figure 4). A realistic spatial trend can also be seen in the RF temperature predictions whereby maximum water temperatures were located in the upper tributaries while minimum water temperatures were highest at the Bay mouth (Figure 4).

RF model predictions for salinity were generally lower than corresponding values from the CBOFS in the mainstem of the Bay and higher than corresponding CBOFS values in the tributaries (Figure 5). CBOFS surface water salinity estimates are known to be saltier than observations [7, 10] partly explaining our saltier data-driven results. The salinity model was able to reproduce realistic temporal dynamics whereby the Bay is generally "fresher" in the Spring season (April-June) and "saltier" in Fall and Winter seasons (Figure 4).

4   Discussion

Our geographically aware approach was able to accurately reproduce realistic spatial and temporal patterns of Bay salinity and temperature. Model errors were generally lower than prior data-driven efforts and as good or better than physics-based models [10]. In addition to improved accuracy, our approach provides the ability to generate rapid bay-wide prediction surfaces in seconds and it is easily updatable insofar as saved models can be updated as new data become available. Our models are lightweight with minimal internal complexity and have been extensively tested against a uniquely large observational dataset.

Although salinity and temperature models were accurate, our turbidity model showed relatively poor performance. One reason for this may have been because our observational data were reported as Nephelometric Turbidity Units (NTUs) which are not standardized. As a result, they cannot be reliably compared across different instruments. A better turbidity model would likely result from using total suspended matter (TSM) data measured in a standardized unit such as mg/l [25]. Another reason the turbidity model performed poorly may be that turbidity is simply difficult to predict. The event-driven nature of turbidity may mean that its extreme spikes may occur over time scales which are too short to be captured by daily reflectance data.

Across all models, MODIS band 8 was the most important reflectance feature. This result may be partly explained by the fact that Band 8 is the shortest wavelength (405-420nm) included in our study and is likely related to the concentration of Colored Dissolved Organic Matter (CDOM) the absorption of which declines exponentially with increasing wavelength under simulated and laboratory conditions [26, 27]. CDOM itself is strongly related to salinity [27, 10, 11] and thus may be related to other properties such as temperature and turbidity which also vary spatially with the upstream-downstream environmental gradient of estuaries. In field conditions and in remote sensing applications, the absoption spectra of CDOM and the shape of the CDOM-salinity relationship is highly variable from region-to-region as a function of local factors such as bottom reflectance and phytoplankton community composition [27]. Although band 8 was not universally the most important feature in prior remote sensing studies of estuarine water properties, the general importance of short wavelength bands is a result shared by prior studies on estuarine salinity [11].

4.1   Spatial patterning

Temperature and salinity models produced realistic predictions insofar as they were characterized by low salinity in the upper tributaries and high minimum water temperatures at the Bay mouth. This indicates that the model captures both freshwater discharge from tributaries as well as the moderating influence of ocean water in areas near the Bay mouth [28].

Our comprehensive dataset includes greater spatial coverage than prior efforts that were largely trained on limited (and relatively high-salinity) buoy or in-situ cruise data located in the mainstem of the Bay [10, 29, 25]. This coverage allows for a more comprehensive assessment of the spatial uncertainty of our model. Predicted areas of low salinity coming from our model likely have particularly low associated uncertainties because the upper tributaries were well-represented in our observational dataset (Figure S3). Because CBOFS salinity data was fresher than our RF estimates in the upper tributaries, this suggests that CBOFS data for these areas are likely true underestimates and not a deficiency of the RF model predictions.

In contrast to the upper tributaries, an area of maximum uncertainty is likely the mainstem and eastern Bay across from the mouth of the Potomac (Tangier Sound). This appears to be the collision point of the

marine plume emanating from the Bay mouth with the discharge of the Potomac River. The hydrodynamics here seem to be complex given that the CBOFS has difficulty resolving spatial patterning in this area [7]. This was an area where we found limited observational data (Figure S3). Improvements to data-driven modeling in this area may be possible with the inclusion of more observational data.

4.2 Temporal patterning

The salinity model was able to reproduce realistic temporal dynamics whereby the Bay is generally "fresher" in the Spring season (April-June) and "saltier" in Fall and Winter seasons. However, because the salinity model was not able to reproduce the highest salinity values Fall and Winter estimates may be biased low. For the temperature model, we were unable to recover low temperature observations less than 0 deg C. These results are consistent with prior research showing winter had the highest seasonal error [29]. One explanation for these high winter error rates is the fact that winter in the Bay is generally characterized by strong vertical mixing due to high winds [30, 1] making "true" patterns more dynamic and more difficult to resolve.

4.3 Future research

One shortcoming of our work which may be a fruitful area of future research is the lack of realistic small-scale features evident in the physics based output of the CBOFS (Figure 5). It may be possible to resolve small-scale features such as eddies and other circulation features with the use of convolutional neural networks (CNNs). A key difference between our approach and that of CNNs is that our approach essentially fits each observation in isolation and does not exploit the "spatial locality" (i.e. the "neighborhood") surrounding observations as is the case with CNNs [31]. The pipeline for training a CNN would bring some practical difficulties regarding the collection and storage of MODIS data. Unlike our RF approach, it would require obtaining entire multiband images for the Bay rather than simply the values that correspond with observational data.

Another potential area of future research is resolving water properties at depths beyond the surface. A key challenge in this area is lack of observational data at multiple locations. Depth arrayed measurements are typically only associated with buoy deployments of which there are few in the Bay [10]. It may be possible to supplement these buoy observations with data from physics-based models such as CBOFS which are fully three-dimensional. The results of such an effort can be regarded more so as an emulator than a standalone model.

5 Conclusion

We have demonstrated an approach to produce realistic estimates of water properties in Chesapeake Bay using remote sensing data. Salinity and temperature model performance was as good or better than prior efforts. Salinity model performance was particularly realistic in low salinity tributary areas. Much of this realism comes from our FWI routines which encourage physical consistency in upstream-downstream spatial patterning. We saw additional performance gains owing tothe fact that our training dataset had spatial coverage in the upper tributaries beyond the mainstem of the Bay. Estimates produced by our model have a high potential to inform calibration and evaluation of Earth Systems Models as well as in other applications such as habitat suitability and sea level rise models.

6 Acknowledgements

References

[1] Jiangtao Xu et al. "Climate Forcing and Salinity Variability in Chesapeake Bay, USA". In: *Estuaries and Coasts* 35.1 (Jan. 2012), pp. 237–261. ISSN: 1559-2723, 1559-2731. DOI: 10.1007/s12237-011-9423-5.

[2] James D. Hagy et al. "Hypoxia in Chesapeake Bay, 1950–2001: Long-term Change in Relation to Nutrient Loading and River Flow". In: *Estuaries* 27.4 (Aug. 2004), pp. 634–658. ISSN: 0160-8347. DOI: 10.1007/BF02907650.

[3] Raymond G. Najjar et al. "Potential Climate-Change Impacts on the Chesapeake Bay". In: *Estuarine, Coastal and Shelf Science* 86.1 (Jan. 2010), pp. 1–20. ISSN: 02727714. DOI: 10.1016/j.ecss.2009.09.026.

[4] Isaac D. Irby et al. "The Competing Impacts of Climate Change and Nutrient Reductions on Dissolved Oxygen in Chesapeake Bay". In: *Biogeosciences* 15.9 (May 2018), pp. 2649–2668. ISSN: 1726-4189. DOI: 10.5194/bg-15-2649-2018.

[5] Raleigh R. Hood et al. "The Chesapeake Bay Program Modeling System: Overview and Recommendations for Future Development". In: *Ecological Modelling* 456 (Sept. 2021), p. 109635. ISSN: 03043800. DOI: 10.1016/j.ecolmodel.2021.109635.

[6] Alexander G. Fore et al. "Combined Active/Passive Retrievals of Ocean Vector Wind and Sea Surface Salinity With SMAP". In: *IEEE Transactions on Geoscience and Remote Sensing* 54.12 (Dec. 2016), pp. 7396–7404. ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2016.2601486.

[7] Lyon Werner John Lanerolle, Richard C Patchen, and Frank Aikman. "The Second Generation Chesapeake Bay Operational Forecast System (CBOFS2) Model Development and Skill Assessment". In: (2011).

[8] Jean-Christophe Golaz et al. *The DOE E3SM Model Version 2: Overview of the Physical Model and Initial Model Evaluation.* Preprint. Climatology (Global Change), Aug. 2022. DOI: 10.1002/essoar.10511174.2.

[9] Donghui Xu et al. "Development of Land-River Two-Way Hydrologic Coupling for Floodplain Inundation in the Energy Exascale Earth System Model". In: *Journal of Advances in Modeling Earth Systems* 14.8 (Aug. 2022). ISSN: 1942-2466, 1942-2466. DOI: 10.1029/2021MS002772.

[10] Ronald L. Vogel and Christopher W. Brown. "Assessing Satellite Sea Surface Salinity from Ocean Color Radiometric Measurements for Coastal Hydrodynamic Model Data Assimilation". In: *Journal of Applied Remote Sensing* 10.3 (July 2016), p. 036003. ISSN: 1931-3195. DOI: 10.1117/1.JRS.10.036003.

[11] Jiao Wang and Zhiqiang Deng. "Development of a MODIS Data Based Algorithm for Retrieving Nearshore Sea Surface Salinity along the Northern Gulf of Mexico Coast". In: *International Journal of Remote Sensing* 39.11 (June 2018), pp. 3497–3511. ISSN: 0143-1161, 1366-5901. DOI: 10.1080/01431161.2018.1445880.

[12] Rebecca R. Murphy, Frank C. Curriero, and William P. Ball. "Comparison of Spatial Interpolation Methods for Water Quality Evaluation in the Chesapeake Bay". In: *Journal of Environmental Engineering* 136.2 (Feb. 2010), pp. 160–171. ISSN: 0733-9372, 1943-7870. DOI: 10.1061/(ASCE)EE.1943-7870.0000121.

[13] Erin A. Urquhart et al. "Geospatial Interpolation of MODIS-derived Salinity and Temperature in the Chesapeake Bay". In: *Remote Sensing of Environment* 135 (Aug. 2013), pp. 167–177. ISSN: 00344257. DOI: 10.1016/j.rse.2013.03.034.

[14] Anantha M. Prasad, Louis R. Iverson, and Andy Liaw. "Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction". In: *Ecosystems* 9.2 (Mar. 2006), pp. 181–199. ISSN: 1432-9840, 1435-0629. DOI: 10.1007/s10021-005-0054-1.

[15] Fabian Pedregosa et al. "Scikit-Learn: Machine Learning in Python". In: *MACHINE LEARNING IN PYTHON* (2011), p. 6.

[16] Erin A. Urquhart et al. "Remotely Sensed Estimates of Surface Salinity in the Chesapeake Bay: A Statistical Approach". In: *Remote Sensing of Environment* 123 (Aug. 2012), pp. 522–531. ISSN: 00344257. DOI: 10.1016/j.rse.2012.04.008.

[17] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17.3 (Mar. 2020), pp. 261–272. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-019-0686-2.

[18] USGS. "USGS Water Data for the Nation: US Geological Survey National Water Information System Database". In: (2022).

[19] Laurie S. Little, Don Edwards, and Dwayne E. Porter. "Kriging in Estuaries: As the Crow Flies, or as the Fish Swims?" In: *Journal of Experimental Marine Biology and Ecology* 213.1 (June 1997), pp. 1–11. ISSN: 00220981. DOI: 10.1016/S0022-0981(97)00006-3.

[20] Stéfan van der Walt et al. "Scikit-Image: Image Processing in Python". In: *PeerJ* 2 (June 2014), e453. ISSN: 2167-8359. DOI: 10.7717/peerj.453.

[21] Eric Vermote and Robert Wolfe. *MOD09GA MODIS/Terra Surface Reflectance Daily L2G Global 1kmand 500m SIN Grid V006*. 2015. DOI: 10.5067/MODIS/MOD09GA.006.

[22] Jemma Stachelek, Sofia Avendano, and Jon Schwenk. *DOE-ICoM/Geowq-Chesapeake*. Zenodo. Nov. 2022. DOI: 10.5281/zenodo.7332558.

[23] Jemma Stachelek. *Salinity, Temperature, and Turbidity Observations during 2000-2021 for Chesapeake Bay*. 2022. DOI: 10.6084/m9.figshare.21578898.

[24] Alexander F. Shchepetkin and James C. McWilliams. "The Regional Oceanic Modeling System (ROMS): A Split-Explicit, Free-Surface, Topography-Following-Coordinate Oceanic Model". In: *Ocean Modelling* 9.4 (Jan. 2005), pp. 347–404. ISSN: 14635003. DOI: 10.1016/j.ocemod.2004.08.002.

[25] Michael. Ondrusek et al. "The Development of a New Optical Total Suspended Matter Algorithm for the Chesapeake Bay". In: *Remote Sensing of Environment* 119 (Apr. 2012), pp. 243–254. ISSN: 00344257. DOI: 10.1016/j.rse.2011.12.018.

[26] Steven A. Loiselle et al. "The Optical Characterization of Chromophoric Dissolved Organic Matter Using Wavelength Distribution of Absorption Spectral Slopes". In: *Limnology and Oceanography* 54.2 (Mar. 2009), pp. 590–597. ISSN: 00243590. DOI: 10.4319/lo.2009.54.2.0590.

[27] Shubha Sathyendranath et al. "Remote Sensing of Ocean Colour in Coastal, and Other Optically-Complex, Waters." In: (2000).

[28] Haiyong Ding and Andrew J. Elmore. "Spatio-Temporal Patterns in Water Surface Temperature from Landsat Time Series Data in the Chesapeake Bay, U.S.A." In: *Remote Sensing of Environment* 168 (Oct. 2015), pp. 335–348. ISSN: 00344257. DOI: 10.1016/j.rse.2015.07.009.

[29] Erick F. Geiger et al. "Satellite-Derived Coastal Ocean and Estuarine Salinity in the Mid-Atlantic". In: *Continental Shelf Research* 63 (July 2013), S235–S242. ISSN: 02784343. DOI: 10.1016/j.csr.2011.12.001.

[30] SeungHyun Son and Menghua Wang. "Water Properties in Chesapeake Bay from MODIS-Aqua Measurements". In: *Remote Sensing of Environment* 123 (Aug. 2012), pp. 163–174. ISSN: 00344257. DOI: 10.1016/j.rse.2012.03.009.

[31] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. "Convolutional Networks". In: *Deep Learning*. Vol. 2016. MIT Press Cambridge, MA, USA, 2016, pp. 330–372.

Biographies

Jemma Stachelek is a Limnologist and Aquatic Earth Scientist (PhD, Msc, BSc) in the Division of Earth and Environmental Sciences at Los Alamos National Laboratory. Her research focuses on the

applications of spatially intensive geocomputation to water and food security problems particularly those related to eutrophication and nutrient pollution.

Sofia Avendaño is a hydrogeologist with a Master's from New Mexico Institute of Mining and Technology. She has worked on a wide variety of projects including fracture flow, surface flow and hydrodynamics. Her scientific passion is modeling fluid flow and sediment transport.

Jon Schwenk is a River and Data Scientist (PhD) in the Division of Earth and Environmental Sciences at Los Alamos National Laboratory. One of his primary research themes is the use of "big data" as it relates to water to drive models addressing pressing water security issues.

Figure Captions

Figure 1: Diagram of the data processing, model fitting, tuning, and prediction pipeline.

Figure 2: Out of sample test set performance for salinity, turbidity, and temperature models respectively. Hatching indicates the location of 80% of the data points.

Figure 3: Random Forest importance plot for salinity, temperature, and turbidity models. Note that all features may not be present in all models if they were dropped as a result of recursive feature elimination.

Figure 4: Seasonal Random Forest prediction results for temperature and salinity.

Figure 5: Comparison between Random Forest salinity prediction results and a CBOFS snapshot for Sept, 4, 2022.