Analyzing sea-level change on the east coast with spatiotemporally correlated data

David W. Coats, Candace Berrett, William F. Christensen, Nathan Sandholtz

Abstract

Increasing rates in sea level rise imply drastic consequences for U.S. coastal populations, infrastructure, ecological systems, and natural resources in the coming decades. These direct impacts will lead to negative repercussions in public health, biodiversity, tourism, and other aspects of the global economy. Using hourly tide readings from the past 30 years at 38 gauges along the east coast, we wish to develop a model that will allow us to analyze the trends in this type of data and to accurately and precisely predict sea level change along the east coast. The model developed is an iterative generalized additive model that will use spatial and temporal dependence between gauges and across time, allowing us to predict sea level change all along the east coast, not only at the stations for which we have data. This generalized additive model includes a linear term, a seasonal trend term fit with B-splines, and a term accounting for additional spatial variance with latent factors estimated by confirmatory factor analysis.

Introduction

The Intergovernmental Panel on Climate Change (IPCC) estimates that global sea levels are currently rising at an average rate of 3 millimeters per year and this rate is expected to increase over the coming century (Solomon et al., 2007). This increase in the rate of sea level rise could lead to changes that will affect many aspects of daily life and the global economy, thus accurate predictions and thorough understanding of the trends in this process are vital for preparation for these changes.

The data set used in this study comes from the National Oceanic and Atmospheric Administration (NOAA) and was compiled by researchers at climate control. This data set consists of tide gauge readings taken hourly from 1979 to 2009 from 38 stations along the east coast of the United States. Tide gauges are instruments deployed at coastal sites around the world that directly measure sea level as compared to a determined base level. Figure 1 shows an example of one of these gauges. For the 38 tide gauges in this study, sea level is measured as deviation from the mean high water level for that station over a 19 year epoch (1983 - 2001).

http://www.oco.noaa.gov/tideGauges.html

Figure 1: Example of a tide gauge from NOAA (National Oceanic and Atmospheric Administration)

The 38 stations where these tide gauges are located range in location from Bar Pilots Dock–St. Johns River in Florida to East Port–Passamaquoddy Bay in Maine. Figure 2 shows the locations of the 38 stations used in this analysis. At some of these stations there are many missing observations due either to malfunctions in the tide gauge or because at that time there was no gauge in that location. These missing values lead to complications in modeling and predicting, but these issues will be addressed later on.

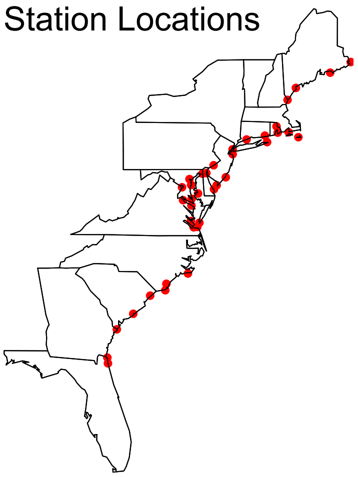

Figure 2: Locations of the 38 tide gauges along the east coast

Due to factors dealing with global location, sea level change is not constant across space in that it is different depending on location. An important observation when considering spatial data such as this is that sites that are closer together are more likely to be closely related than sites that are further apart. Our model will take into account these spatial correlations, and this will eventually allow us to predict along the coast because of the spatial relationship between these sites and sea level.

Exploratory Analysis: Exposing Trends

To understand the temporal trends in our data, we identify patterns seen in sea levels at each station over time, and model these trends for individual sites. After identifying and exploring these trends in individual sites we will seek to combine these trends in an iterative spatial generalized additive model that will use correlation between sites in order to predict at any location along the east coast. Because ocean tides are greatly influenced by the moon and its cycles, we average the 30 years of data by lunar months which are approximately 28 days long, resulting in 371 lunar month averages for each station. A general linear trend can be seen in these lunar month averages over time in all of the stations, but there are other trends in addition to a simple linear relationship. Figure 3 shows the lunar month averages across time with a simple linear fit plotted on top (the blue line). Note the missing data between 1996 and 1998.

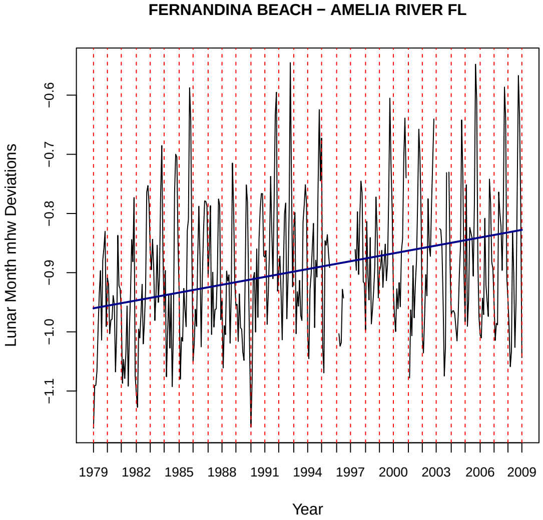

Figure 3: The linear trend in lunar month average across time seen at a single station

There is a visible linear trend in the lunar month averages, but there appears to be a cyclical effect within years that could be indicative of a seasonal trend. After accounting for the simple linear trend at each station, we compute the residuals and see a definite pattern. The residuals for the southern most stations have what we will call an M-curve, but as we move north, station by station this M-shape appears to flatten out into flatter, more unimodal curves suggesting that this M-curve effect changes across space, illustrated in Figure 3.2. We fit these curves individually for each station with B-splines. A B-spline with k knots splits the covariate space into $k+1$ regions fitting a cubic polynomial to the data in each region. The spline is constrained so that the polynomials are differentiable at the knot points resulting in a smooth non-linear fit to the data (James et al., 2013). Figure 4 shows the residuals for a north station and a more southern station fit with B-splines with 6 knots.

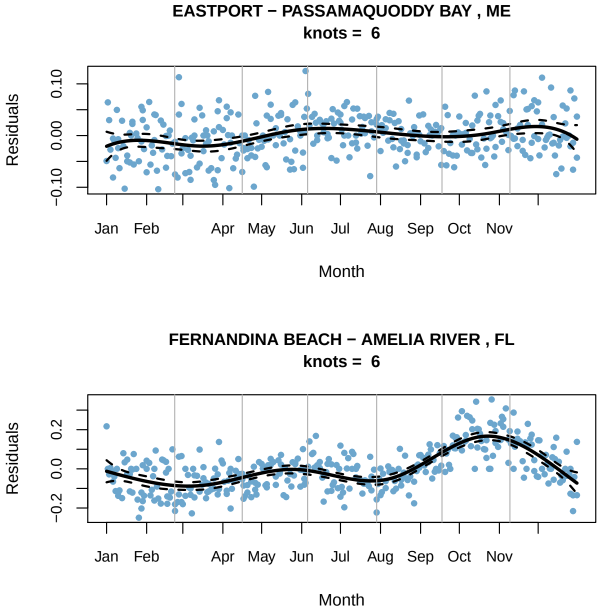

Figure 4: The seasonal trends for a northern and southern station fit with B-splines with 6 knots

These splines appear to be good fits to the residuals, and because the seasonal trend seems to have spatial correlation, we will attempt to fit spatial terms to these splines so that we can model the relationship of this trend with spatial location. Now that we have uncovered some important trends in the data, we develop a model that iteratively fits these trends and uses spatial correlation to explain the relationships between sites, allowing for more extensive predictive power and better understanding of sea-level changes across space and time.

Iteritive Generalized Additive Model

Generalized additive models are a generalization of linear models in which the predictions depend on smooth functions of the covariates (Hastie and Tibshirani, 1990). We will model Y_{it}, sea-level change at station i and time t, with an intercept μ_i that will be the overall mean of each station, a linear term, and a spline term that accounts for the seasonal trend. This model will be expressed in the following manner:

$$Y_{it} = \mu_i + t\beta_i + \sum_{j=1}^{k+1} g_j(t^*)\xi_{ij} + \nu_{it} \qquad \nu_{it} \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

where $t\beta_i$ is time (centered and scaled) multiplied by the coefficient vector β for station i, $g_j(t^*)$ is the polynomial fit to the j^{th} of $(k+1)$ covariate regions where $k=6$ is the number of knots in our B-spline, t^* is the day of the year, ξ_{ij} is the coefficient fit to the j^{th} covariate region for the i^{th} station, and ν_{it} is the unexplained variance in sea-level change at station i and time t.

We refer to our model as an iterative generalized additive model because we will fit the terms iteratively. We will fit the linear term to the residuals of the model containing only μ_i, then we will fit the spline term to the residuals of the model with the intercept plus the linear term. Note that $\hat{\mu}_i$ will simply be calculated as the mean of all observed values for station i. We fit the model in this manner in order to address the problem we have with missing observations. For the first iteration, the missing observations were all replaced with the overall mean of the station to which they belong as a beginning value. For each step in fitting the model we predict Y_{it} and then replace the formerly missing values with $\hat{Y}_{it}$ iteratively until the predictions of these observations and the coefficients of the model converge to specific values. Convergence is met when the $\hat{Y}_{it}$s for the originally missing observations change by less than tolerance level τ for the current iteration compared to the last. Fitting the model this way allows us to update the values of the missing observations based on the trends in the data step by step, resulting in better estimates for these values at every step and a better final estimate.

Inclusion of Spatial Correlation in Model

Previously we fit a linear trend and a spline term individually to each station; now we will model the relationship or correlation of these trends from station to station. Understanding how spatial distance affects correlation between stations will allow us to be able to make inference along the coast between our stations. Given the data at our stations and the distances from new locations to our stations, we will be able to predict sea-level changes at these new locations.

A semi-variogram is a function describing the correlation between points that are different distances apart. Semi-variograms can be modeled with different spatial correlation structures that behave differently depending on how the data is spatially correlated (Waller and Gotway, 2004). The Matern, exponential, Gaussian, and spherical functions are examples of spatial correlation structures; by exploring the fits of these different functions to the residuals left over after taking out the linear trend, we decide that the spherical function is the best fit. Assuming the spherical semi-variogram is a good fit to the semi-variogram of our data, the semi-variance between two observations with distance $\ell < \phi$ between them is

$$\gamma(\ell) = (1-c_o)\frac{3\ell}{2\phi} + \frac{1}{2}\left(\frac{\ell}{\phi}\right)^3 \tag{2}$$

for all observations for which $\ell > 0$ where ϕ is the range over which the correlations will be nonzero, and c_o is the nugget. The range ϕ refers to the distance at which the semi-variogram appears to level out because points with distances greater than ϕ are not correlated. For reference to the terminology used in modeling semi-variograms, the semi-variance at distance ϕ is σ^2, referred to as the sill, and the partial sill c_e is $\sigma^2 - c_o$; the nugget c_o is the semi-variance at distance 0, meaning that if c_o is non-zero, that there is variance among points that are very close together, indicating randomness or possibly underlying trends in the residuals. Figure 5 depicts a semi-variogram of a set of residuals fit with the spherical function with the estimated nugget, partial sill, and sill denoted.

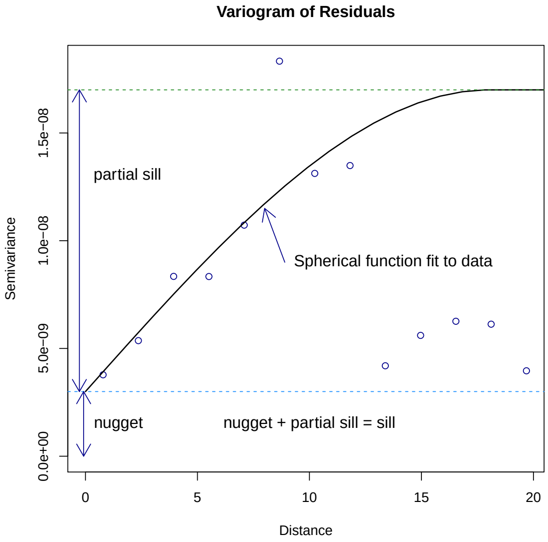

Figure 5: Variogram of the residuals from the linear model fit with spherical correlation function

In order to estimate the coefficient β_i for the linear term, we first fit $Y_{it} - \mu_i = t\beta_i^*$, where β_i^* is the coefficient fit to centered time t for station i. We then fit $\beta_i^* = \alpha_{00} + x_i\alpha_{01} + x_i^2\alpha_{02}$ using ordinary least squares with x_i being the coastal distance of site i. We estimate $\hat{\boldsymbol{\alpha}}_0 = (\hat{\alpha}_{00}, \hat{\alpha}_{01}, \hat{\alpha}_{02}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\beta}^*$ with $\mathbf{X}$ being a matrix with a column of ones, a column of $x_i s$, and a column of $x_i^2 s$. This results in an estimate of $\boldsymbol{\beta}^*$ in $\boldsymbol{\beta}^{**} = \mathbf{X}\hat{\boldsymbol{\alpha}}_0$.

We then calculate the residuals $r_i = \beta_i^* - \beta_i^{**}$ and estimate the spatial covariance of r_i and r_j denoted c_β, r_j being the the residual for station j, separated by distance ℓ with

$$\hat{c}_\beta(\hat{\boldsymbol{\theta}}) = \sigma^2 - \gamma(\ell, \hat{\boldsymbol{\theta}}) \tag{3}$$

where $\hat{\boldsymbol{\theta}}$ is the estimate of the semi-variogram parameters. We then fit $\beta_i^* = \alpha_0 + x_i\alpha_1 + x_i^2\alpha_2$ using generalized least squares. We estimate $\hat{\boldsymbol{\alpha}} = (\hat{\alpha_0}, \hat{\alpha_1}, \hat{\alpha_2}) = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\boldsymbol{\beta}^*$ with x_i being the coastal distance of site i from the southern most station, and $\mathbf{W}$=$\hat{\boldsymbol{\Sigma}}_\beta(\boldsymbol{\theta})$ being the covariance matrix made up of the covariances $\hat{c}_\beta$. The resulting spatially dependent estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \mathbf{X}\hat{\boldsymbol{\alpha}}$.

The step of fitting the spatial covariance to the residuals is repeated, the new residuals being $r_i^+ = \beta_i^* - \hat{\beta}_i$, until the estimates of $\hat{\boldsymbol{\beta}}$ converge to a value. $\hat{\boldsymbol{\beta}}$ is determined to have converged when each $\hat{\beta}_i$ no longer changes compared to the $\hat{\beta}_i$ computed in the last iteration.

In fitting the data available for this study, no additional iterations were needed to converge to a value of $\hat{\boldsymbol{\beta}}$ as the $\mathbf{W}$ estimated in the second iteration was equal to the $\mathbf{W}$ produced in the first iteration. Using these methods, we have modeled the trend of how the linear effect changes across space while maintaining accurate modeling at individual stations.

Now for the spline term $\sum_{j=1}^{k+1} g_j(t^*)\xi_{ij}$, we model the coefficient vector $\boldsymbol{\xi}_{\cdot j}$ for each of the 7 regions formed by 6 evenly spaced knots in the same manner as we modeled $\boldsymbol{\beta}$ but fit to the residuals of the model including $t\hat{\boldsymbol{\beta}}$, $Y_{it} - \mu_i - t\hat{\beta}_i$. After estimating ξ_{ij}^* and $\mathbf{V}_j = \boldsymbol{\Sigma}_{\xi_j}(\boldsymbol{\theta})$, for the i^{th} station and j^{th} of the 7 regions across covariate space, and solving for $\hat{\boldsymbol{\gamma}}_j = (\hat{\gamma}_{j0}, \hat{\gamma}_{j1}, \hat{\gamma}_{j2}) = (\mathbf{X}'\mathbf{V}_j^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_j^{-1}\boldsymbol{\xi}_j^*$, we have $\xi_{ij} = \gamma_{j0} + x_i\gamma_{j1} + x_i^2\gamma_{j2}$ by generalized least squares resulting in the spatially dependent estimate for $\boldsymbol{\xi}_{\cdot j}$ in $\hat{\boldsymbol{\xi}}_{\cdot j} = \mathbf{X}\hat{\boldsymbol{\gamma}}_j$. Just as with the fitting of $\boldsymbol{\beta}$, we solve for $\hat{\boldsymbol{\xi}}_{\cdot j}$ iteratively, fitting $\mathbf{V}_j$ to the residuals $r_{ij}^+ = \xi_{ij}^* - \hat{\xi}_{ij}$, until the estimate $\hat{\boldsymbol{\xi}}_j$ converges.

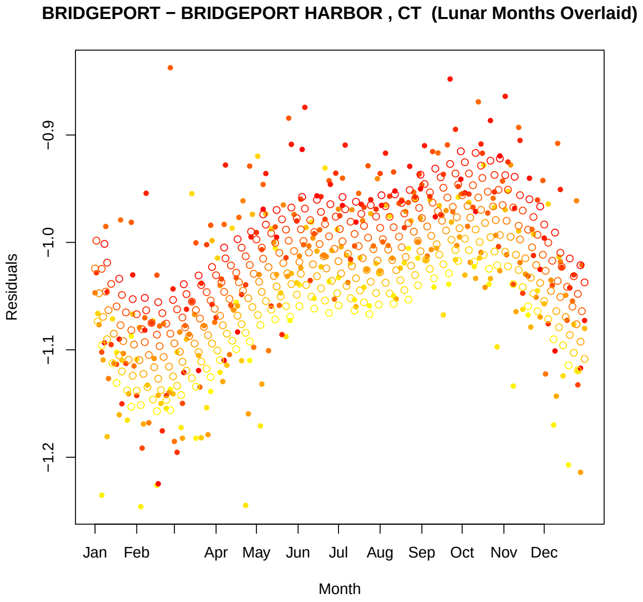

Figure 6: Predicting the residuals from the spatial model including the intercept, linear, and spline terms at a single station by time of the year. The solid points are the observed residuals and the open points are predicted residuals. Yellow represents residuals and predictions for observations from earlier years in the data and red represents observations from later years.

Figure 6 shows for a single station that because of the positive linear trend modeled by $t\beta_i$, the model predicts the M-curves for more recent years to be higher than the M-curves for earlier years. Figure 7 is an example of what the model fit looks like at a single station, the blue being the predicted lunar month average, and black being the lunar month averages from the actual data.

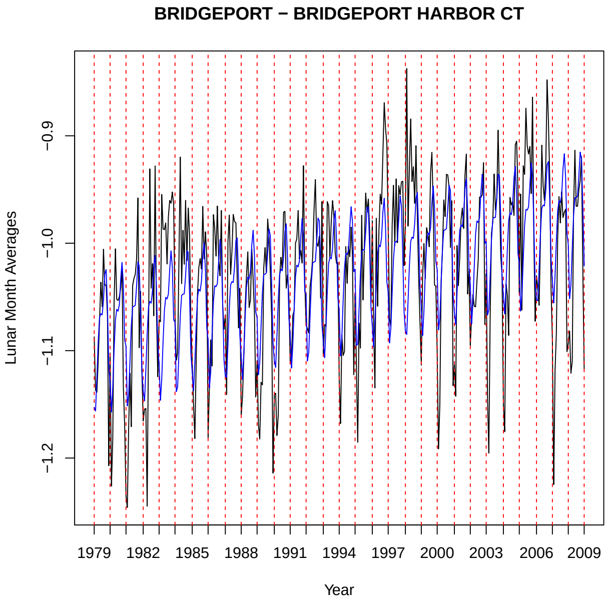

Figure 7: The linear trend in lunar month average across time seen at a single station

To this point in the model fitting process, we have iteratively fit a station mean term, a linear term across time, and a seasonal trend with B-splines. In order to be able to predict anywhere along the coast, not just at the 38 stations, we have modeled the spatial covariance of the data using the spherical theoretical semi-variogram.

Confirmatory Factor Analysis

In an effort to account for latent factors in the data that the current model has not uncovered, we will perform confirmatory factor analysis on the remaining residuals seeking an identified solution to the factor analysis model

$$\begin{aligned}
z_{1t} &= \lambda_{11}f_{1t} + \lambda_{12}f_{2t} + \cdots + \lambda_{13}f_{3t} + \epsilon_{1t} \\
z_{2t} &= \lambda_{21}f_{1t} + \lambda_{22}f_{2t} + \cdots + \lambda_{23}f_{3t} + \epsilon_{2t} \\
&\vdots \\
z_{38t} &= \lambda_{381}f_{1t} + \lambda_{382}f_{2t} + \cdots + \lambda_{383}f_{3t} + \epsilon_{38t}
\end{aligned} \tag{4}$$

where the factor loading λ_{ij} quantifies the influence of the factor $f_{j\cdot}$ on the observed variable $z_{i\cdot}$, which in this model represents the residuals at station i. Further, z_{it} represents the i^{th} station at time t, f_{jt} is the j^{th} factor for the the residuals at time t, defined $\mathbf{z}_t$, and ϵ_{it} represents the specific factor (or error) component for z_{it}. We choose to constrain the number of latent factors m to be 3 because exploratory analysis indicates the majority of variance in the residuals can be explained by the first 3 eigenvalues of the residual matrix. In matrix form, we write the model

$$\mathbf{z}_t = \mathbf{\Lambda f}_t + \boldsymbol{\epsilon}_t. \tag{5}$$

The factor vector $\mathbf{f}_t = (f_1, \ldots, f_m)'$ for time t has mean $\mathbf{0}$ and covariance matrix $\mathbf{\Phi}$ and the error vector $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{pt})'$ has mean $\mathbf{0}$ and diagonal covariance matrix $\mathbf{\Psi}$. The model will be constrained by fixing a 3×3 sub-matrix of $\mathbf{\Lambda}$ to be equal to the identity matrix, thus defining each of 3 variables to be equal to a factor plus error.

Using the parameter estimates $\hat{\mathbf{\Lambda}}$, $\hat{\mathbf{\Phi}}$, and $\hat{\mathbf{\Psi}}$ in the *regression method* formula

$$\hat{\boldsymbol{f}}_{\boldsymbol{t}} = \hat{\mathbf{\Phi}}\hat{\mathbf{\Lambda}}'(\hat{\mathbf{\Lambda}}\hat{\mathbf{\Phi}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}})^{-1}(\mathbf{z}_t - \bar{\mathbf{z}}) \tag{6}$$

we can form factor score estimates for time t. We can now estimate $\mathbf{z}_t$ with

$$\hat{z}_t = \hat{\mathbf{\Lambda}}\hat{\boldsymbol{f}}_t \tag{7}$$

which results in the matrix $\hat{\mathbf{Z}} = (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_{371})'$ that may be added to the GAM in (1) as another term in order to account for the latent factors that persist in the residuals of the previously fit model. In order to be able to predict factor loadings for locations between stations, we propose fitting smooth functions to the 3 factor loading vectors $\hat{\boldsymbol{\lambda}}_{\cdot 1}, \hat{\boldsymbol{\lambda}}_{\cdot 2}$, and $\hat{\boldsymbol{\lambda}}_{\cdot 3}$.

Results

In order to better evaluate the predictive capabilities of the proposed model, the data was split into training and test sets in an effort to cross-validate the results. The 4 stations with the most missing data as well as 2 randomly chosen stations were taken from our data set as the test set, and the remaining 32 stations served as our training set on which the model was fit. We then predicted lunar month averages for mean high-water deviation at each of the 6 test stations using the model developed on the training stations. The model up to fitting the confirmatory factor analysis on the training set produced satisfactory results similar to those seen on the model fit to the entire data set.

In this chapter we discuss the results of the confirmatory factor analysis fitted to the residuals of the model that includes the overall mean, linear, and seasonal trend terms. We also discuss the resulting predictions for stations in the training set. We will then discuss the prediction procedure for the test set and show the resulting predictions.

Confirmatory Factor Analysis: Results

As previously described, confirmatory factor analysis was performed on the residual matrix $\mathbf{Z}$, the residuals for the 32 training stations $(z_{1\cdot}, ..., z_{32\cdot})$ being the multivariate response variables, and z_{it} being a residual for the ith station at time t where t goes from 1 to 371 lunar months starting in 1979. We fit the confirmatory factor analysis using 3 factors, constraining each of 3 stations to be equal to one of the 3 factors in order to reduce rotational ambiguity. During exploratory analysis we chose 3 stations that appeared to load high on one of the 3 factors. Assuming station i is one of these 3 stations, we set the λ_{ij} for station i and factor j to 1 and the other two λ_{ij}'s for that station to 0.

The resulting estimates for the factor loadings $\hat{\lambda}_{ij}$ indicate that there is a latent "northness" variable in that stations towards the north appear to load high on factor 1, more central stations load high on factor 2, and the southern stations load high on factor 3. Figure 8 depicts the factor loadings by coastal distance, coastal distance being the distance along the coast from the southern most station.

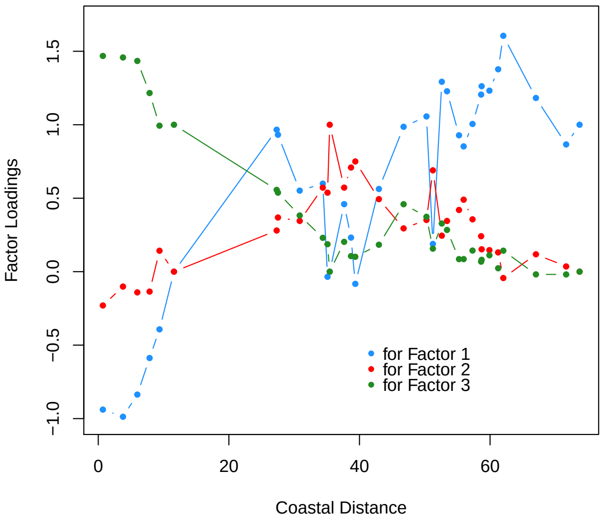

Figure 8: Factor loadings for the 3 factors by coastal distance from the most southern station

Now in an effort to predict the remaining residuals z_{ij}, we estimate the 371 $\hat{\boldsymbol{f}}_t$'s by (6). Figure 9 shows the estimates of the 3 factors across t. Depending on a station's factor loadings, its temporal trend is more or less influenced by a combination of the 3 latent factors.

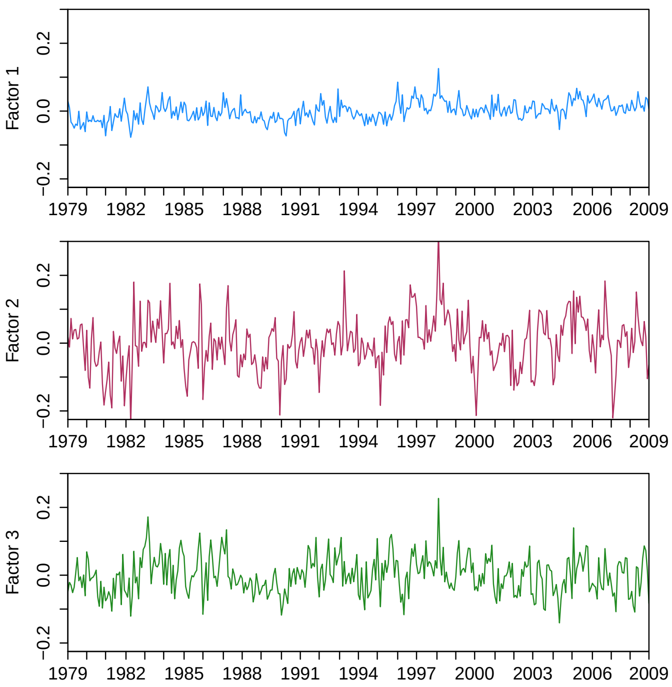

Figure 9: Estimated values of the 3 latent factors across the time

We now calculate the estimated residual matrix $\hat{\mathbf{Z}} = \hat{\mathbf{\Lambda}}\hat{\boldsymbol{F}}$ and add $\hat{\mathbf{Z}}$ to the model expressed in (1) resulting in

$$Y_{it} = \mu_i + t\beta_i + \sum_{j=1}^{k+1} g_j(t^*)\xi_{ij} + (\hat{\mathbf{\Lambda}}\hat{\boldsymbol{F}})_{it} + \epsilon_{it} \qquad \epsilon_{it} \sim \mathcal{N}(0, \sigma^2) \tag{8}$$

where $\hat{\mathbf{\Lambda}}$ is a 32 by 3 matrix made up of the factor loadings for the 3 factors at each of the 32 training set stations, $\hat{\boldsymbol{F}}$ is a 3 by 371 matrix made up of estimated factor values for the 3 factors at the 371 lunar month times, ϵ_{it} is the error left over after fitting confirmatory factor analysis to the residuals of the model in (1), and all other parameters are as defined in the corresponding section. The resulting model produces visibly better predictions of sea level (mean high water deviations as defined in Introduction) in the training stations than the model without the addition of $\hat{\mathbf{Z}}$.

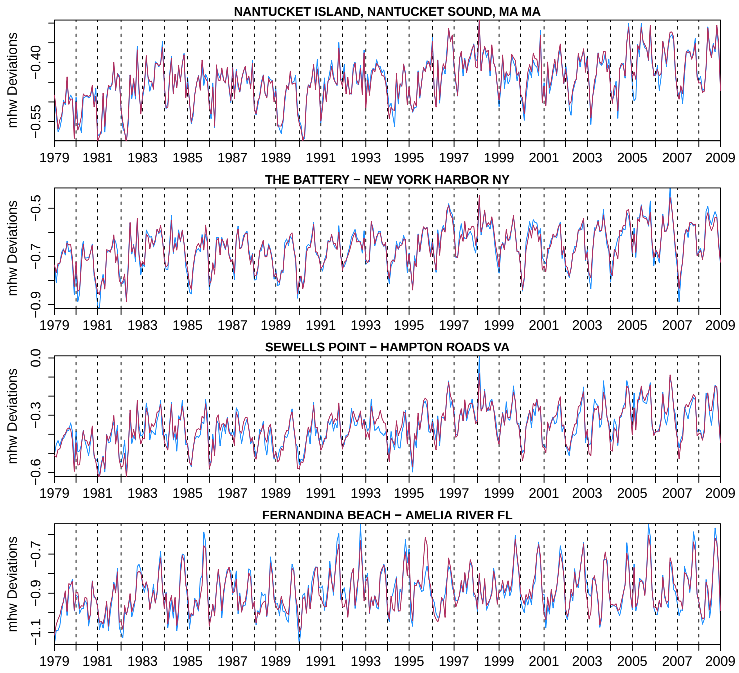

Figure 10: Predictions for 4 selected training set stations, blue lines indicate the original data and red lines are the predicted mhw deviations for that station

In Figure 10 we see that our predicted mean high water (mhw) deviations match the patterns in the data very well for these selected training set stations, as they do in the other 28 training stations. It is not extremely surprising that the addition of the estimated residuals to our model improves prediction, but the interesting question is how will the predictions for the test set stations be affected? We now focus on the prediction of mhw deviation at stations for which we only have information on spatial location and no relevant tide gauge data.

Predicting along the coast

Because our model was developed using spatial covariance structures, we can estimate $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_{\cdot j}$ for any location along the east coast given longitude and latitude as well as coastal distance, which can be calculated given longitude and latitude. In our model fit to the training data we used the site means $\bar{\boldsymbol{\mu}}$ as an estimate for $\boldsymbol{\mu}$ because we believe that using these estimates at the sites we know will give the model strength when predicting $\boldsymbol{\mu}$ between stations.

To estimate $\boldsymbol{\mu}$ for the test sites, we fit a linear model to the training set station means with coastal distance and coastal distance squared as covariates as we did for modeling $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_{\cdot j}$ resulting in the estimate $\boldsymbol{\mu}^*$. We then calculated the residuals $\boldsymbol{\mu}^*$ - $\bar{\boldsymbol{\mu}}$ and calculated the spatial covariance matrix $\mathbf{U} = \hat{\mathbf{\Sigma}}_{\mu}(\hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is the vector of estimated semi-variogram parameters fit to the semi-variogram of the residuals. $\hat{\boldsymbol{\mu}}$

for the training set was then calculated by generalized least squares with $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\alpha}}$ where $\hat{\boldsymbol{\alpha}} = (\hat{\alpha_0}, \hat{\alpha_1}, \hat{\alpha_2}) = (\mathbf{X'U}^{-1}\mathbf{X})^{-1}\mathbf{X'U}^{-1}\bar{\boldsymbol{\mu}}$ with covariance matrix $\mathbf{U}$ and covariate matrix $\mathbf{X}$, the same $\mathbf{X}$ used in estimating $\boldsymbol{\mu}$ with $\boldsymbol{\mu}^*$.

Because we have now constrained $\hat{\mu}_i$ to be spatially correlated to the surrounding $\hat{\mu}_j$'s, we can estimate the overall mean for a new set of locations $\boldsymbol{\mu}_{pred}$ with $\hat{\boldsymbol{\mu}}_{pred}$ using the multivariate normal conditional distribution to get

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{pred} = E(\boldsymbol{\mu}_{pred}|\boldsymbol{\mu}_{train}) = \mathbf{X}_{pred}\hat{\boldsymbol{\alpha}} + \\ \mathbf{U}_{pred,train}\mathbf{U}^{-1}_{train,train}(\bar{\boldsymbol{\mu}}_{train} - \hat{\boldsymbol{\mu}}_{train})\end{aligned} \qquad (9)$$

where the subscript *pred* indicates the subset of that matrix that pertains to the stations to be predicted at, and the subscript *train* indicates the subset of the matrix pertaining to the training set stations. Figure 11 shows the predicted $\hat{\mu}_i$'s along with the actual station means. When we see the mhw deviation predictions for the test set stations in the following pages we will see that some of our predictions are high overall or low overall but seem to fit the linear, seasonal, and ”northness” factor trends well. The quadratic term we used in fitting the linear regression to the station means appears to be a satisfactory fit to the training set, and the spatial covariance appears to place predictions where we would want, but some of our test stations don't seem to follow the patterns seen in the training means quite as well as we would like. The fitting of the overall mean intercept $\boldsymbol{\mu}$ is a part of our model that will require further investigation in an effort to produce better predictions.

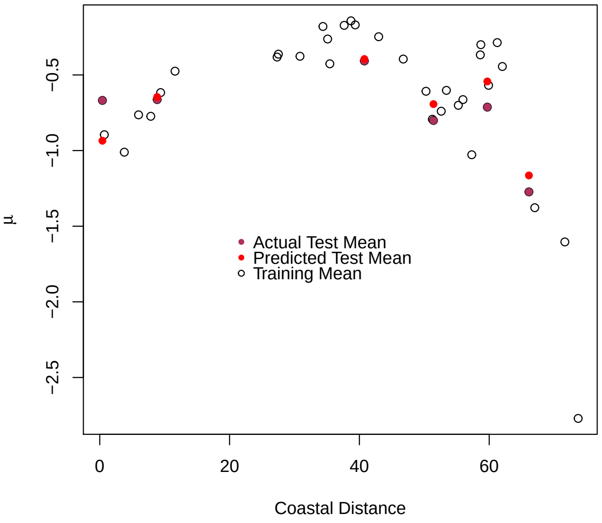

Figure 11: Station means and estimated means for test set stations by coastal distance

Value vectors for $\hat{\boldsymbol{\beta}}_{pred}$ and $\hat{\boldsymbol{\xi}}_{jpred}$ are obtained in the same manner as for $\hat{\boldsymbol{\mu}}_{pred}$ with

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{pred} = E(\boldsymbol{\beta}_{pred}|\boldsymbol{\beta}_{train}) = \mathbf{X}_{pred}\hat{\boldsymbol{\alpha}} + \\ \mathbf{W}_{pred,train}\mathbf{W}^{-1}_{train,train}(\boldsymbol{\beta}^*_{train} - \hat{\boldsymbol{\beta}}_{train})\end{aligned} \qquad (10)$$

and

$$\begin{aligned}\hat{\boldsymbol{\xi}}_{j(pred)} = E(\boldsymbol{\xi}_{j(pred)}|\boldsymbol{\xi}_{j(train)}) = \mathbf{X}_{pred}\hat{\gamma}_j + \\ \mathbf{V}_{j(pred,train)}\mathbf{V}^{-1}_{j(train,train)}(\boldsymbol{\xi}^*_{j(train)} - \hat{\boldsymbol{\xi}}_{j(train)})\end{aligned} \qquad (11)$$

respectively, with all matrices and parameters as defined in (1).

We now have the pieces to form mhw deviation predictions that account for the overall station mean, the simple linear trend, and the seasonal trends seen in the data. In order to estimate $\mathbf{Z}_{pred}$ to account for the “northness” factor in the test stations, and any east coast location at which we wish to predict, we use linear interpolation to get factor loadings $\hat{\boldsymbol{\Lambda}}_{pred}$. We then calculate $\hat{\mathbf{Z}}_{pred} = \hat{\boldsymbol{\Lambda}}_{pred}\hat{\boldsymbol{F}}$ using the same $\hat{\boldsymbol{F}}$ from (8). Now by (8) we add the 4 pieces of our general additive model together for our test set stations and we have predicted mhw deviation lunar month means at sites for which we used only corresponding spatial locations in getting predictions.

In Figure 12 we see that our predictions appear to be very good for Chesapeake City, MD and Southport, NC but shifted vertically compared to the actual data for the other 4 test stations. We are satisfied with the results in that the predictions look very good for all 6 stations barring the bias in estimating the overall mean or intercept term $\boldsymbol{\mu}$. Our predictions appear to estimate well the overall linear trend over time as well as the periodic trends pulled out by CFA, and the seasonal trends within years.

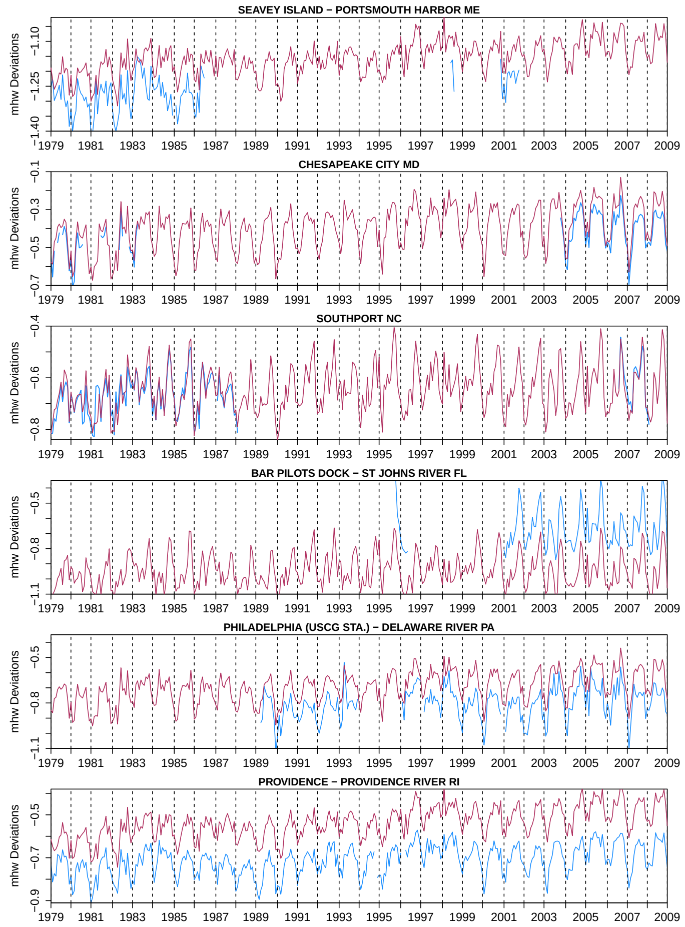

Figure 12: Predictions for the 6 test set stations, blue lines indicate the original data and red lines are the predicted mhw deviations for that station

Conclusions

Due to factors dealing with global location, sea level rise is not constant across space in that it differs by location. An important observation when considering spatial data such as this is that sites that are closer together are more likely to be closely related than sites that are further apart. The model developed in this project takes into account these spatial correlations, and allows us to predict up and down the coast because of the spatial relationship between these sites and sea-level change.

We have seen that there is a general upward trend in mhw(mean high water) deviation across the time span of our data set (1979 to 2009), some stations increasing more drastically than others. We see that generally mhw deviations for stations in the south are increasing more rapidly than those in the north. We have also seen that there is a seasonal trend within years that typically follows a "M-shaped" curve, the steepness of parts of the curve being different by station, but the general shape appearing to flatten out as we go north from the southern stations. After performing confirmatory factor analysis on the residuals left over from modeling the overall mean, linear, and seasonal trends, we see that there are 3 significant latent factors that contribute to what we will call the "northness" factor, stations in the north region loading more on factor 1, central stations loading more heavily on factor 2, and the southern stations loading more on factor 3.

After fitting the previously described model with spatial covariance matrices that correlate parameter estimates for stations depending on the distance between them to a training set, we predicted mhw deviations at 6 stations that we held out as our test set. We are satisfied overall that our model produces reasonable predictions for our test set, noting the need for considerable improvement in predicting the station intercepts. One possible solution that will be looked into is the inclusion of the mean high water mark for each station which may lead to a smoother relationship between stations for the intercepts versus using the overall station means. In addition, when looking at Figure 9 that plots the 3 factors across time, we see a linear trend across time for factor 1. We would expect that this trend would have been accounted for by the linear trend in our model, but this trend seen in factor 1 indicates that by constraining the stations to be spatially correlated we may have forced the northern stations or stations that load high on factor 1 to have smaller linear coefficients.

We propose in future work to fit smoothing functions to the factor loadings in order to get better predictions than using interpolated factor loadings. To this point in the process we do not have a good way to measure uncertainty on our mhw deviation predictions, but we can obtain uncertainty measurements on the parameters $\boldsymbol{\mu}_{pred}, \boldsymbol{\beta}_{pred}$, and $\boldsymbol{\xi}_{j(pred)}$ using the formula for variance given by the multivariate normal conditional distribution, related to how we calculated the point estimates.

This type of model that incorporates spatial correlation between tide gauge stations can be a very effective way to model sea level rise because it allows researchers to predict sea level at locations they do not have data for and model how general trends change along a coast.

References

Arnell, N. W., Livermore, M., Kovats, S., Levy, P., Nicholls, R., Parry, M., and Gaffin, S. (2004), "Climate and socio-economic scenarios for global-scale climate change impacts assessments: Characterising the SRES storylines," *Glob. Environ. Chang.*, 14, 3–20.

Church, J. A., and White, N. J. (2011), "Sea-Level Rise from the Late 19th to the Early 21st Century," *Surv Geophys*, 32, 585–602.

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized additive models* (Vol. 43), CRC Press.

James, G., Witten, D., Hastie, T. J., and Tibshirani, R. J. (2013), *An introduction to statistical learning*, Springer.

Overpeck, J. T., Otto-Bliesner, B., Miller, G. H., Muhs, D. R., Alley, R. B., and Kiehl, J. (2006), "Paleoclimatic evidence for future ice-sheet instability and rapid sea-level rise," *Science*, 311, 1747–50.

Rencher, A. C., and Christensen, W. F. (2012), *Methods of Multivariate Analysis*, John Wiley and Sons, Inc.

Solomon, S., Qin, D., Manning, M. J., Chen, Z., Marquis, M., Averyt, K., Tignor, M. M. B., and Jr, H. L. M. (2007), *Climate Change 2007 The Physical Science Basis: Contribution of Working Group I to the Fourth Assessment Report of the IPCC*, Cambridge Univ Press, New York.

Tebaldi, C., Strauss, B. H., and Zervas, C. E. (2012), "Modelling sea level rise impacts on storm surges along US coasts," *Environmental Research Letters*, 7.

Vermeer, M., and Rahmstorf, S. (2009), "Global sea level linked to global temperature," *Proceedings of the National Academy of Sciences*, 106, 21527–32.

Waller, L. A., and Gotway, C. A. (2004), *Applied Spatial Statistics for Public Health Data*, John Wiley and Sons, Inc.

Wu, S. Y., Najjar, R., and Siewart, J. (2009), "Potential impacts of sea-level rise on the mid and upper Atlantic region of the United States," *Climatic Change*, 95, 121–38.