Florida Storm Surge: Bayesian Gaussian Process Modeling and Computation for Real-World Spatial and Temporal Data

Student: John Mahlon Scott
Supervisor: Hsin-Hsiung Bill Huang, PhD, Associate Professor
We thank National Science Foundation DMS-1924792, DMS-2318925 (Huang)

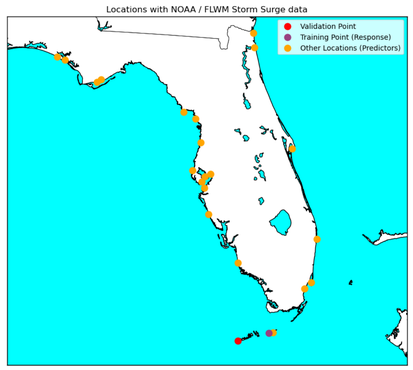
Department of Statistics and Data Science
University of Central Florida, Orlando, FL

Introduction

- Hurricanes and storm surge are a constant threat to the coastlines of Florida. There are many existing ways of modeling storm surges, typically involving the use of detailed bathymetry data and simulating the fluid dynamics. We attempt to create a model for storm surge at locations that have limited bathymetric and historical data.

Dataset

- The data used for this model include full hourly historical storm surge, along with associated wind speed, pressure, and precipitation data for about 25 locations within the state of Florida from the period 1979 - 2014.
- We attempt to find a reasonable bound for the Surge in Key West based on the a Surge model in Key Marathon



Bayesian Spatiotemporal Model

- We will call the surge at the nearest point at which historical surge data is available Z_{ts}, where t is the index for time, and s is the index for space.

$$Z_{ts} = \beta_1^\top X_{t+d_s} + \beta_2^\top X_{t+d_s-1} + \beta_3^\top X_{t+d_s+1} + \alpha_t + \gamma_s + \epsilon_s + \epsilon_t$$

- And the priors are (hyperpriors not shown for brevity, all are uninformative):

$$\begin{aligned}
\beta &= \left[\beta_1^\top,\ \beta_2^\top,\ \beta_3^\top\right]^\top \sim N(\beta_0, \sigma_\beta^2 I_{3p}) \\
\alpha_t &\sim N(0, \Sigma_{l_t}\sigma_{tk}^2) \\
\gamma_s &\sim N(0, \Sigma_{l_s}\sigma_{sk}^2) \\
\epsilon_s &\sim N(0, \sigma_{sn}^2) \\
\epsilon_t &\sim N(0, \sigma_{tn}^2) \\
d_{sp} + \frac{n}{2} &\sim \text{Binom}(n = 200, q_{sp})
\end{aligned}$$

- d accounts for the delay between the various time series used in the model.
- α_t and γ_s create a spatiotemporally varying intercept via gaussian processes which use the RBF kernel.

Posterior Sampling Details

Algorithm 1 Gibbs Sampling for parameters in 1-location Model

Initialize $\vec{d}, \vec{q}, \sigma_{tk}^2, \sigma^2 = \sigma_{tn}^2, \sigma_\beta^2, \vec{\alpha}, \vec{\beta}, \beta_0$
$X \leftarrow [\text{shift}(X, d),\ \text{shift}(X, d-1),\ \text{shift}(X, d+1)]$
Create B and d according to NNGP rules, s.t. $\sigma_{tk}^{-2}\Sigma_{l_t}^{-1} + \sigma^{-2}I = B^\top \text{diag}(d)^{-1}B$

for i in samples **do**

$V \leftarrow \left(\sigma_\beta^{-2} I_{3p} + \sigma^{-2} X^\top X\right)^{-1}$
$m \leftarrow V(\sigma_\beta^{-2} I_{3p}\beta_0 + \sigma^{-2} X^\top (Z - \alpha))$
Sample $\vec{\beta}$ from $N(m, V)$

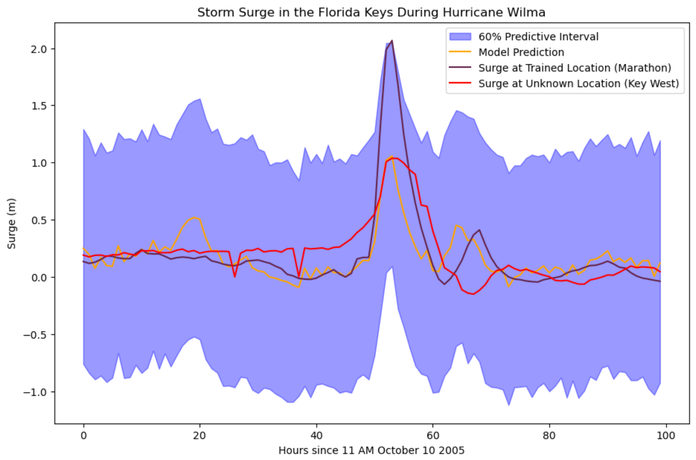
$a \leftarrow 0.01 + (n/2)$
$b \leftarrow 0.01 + 0.5(z - X\beta - a)^\top (z - X\beta - a)$
Sample σ^2 from $IGamma(a, b)$

Sample each $d[p]$ by Metropolis-Hastings
$X \leftarrow [\text{shift}(X, d),\ \text{shift}(X, d-1),\ \text{shift}(X, d+1)]$
$a \leftarrow 3 + \vec{d} + (n/2)$
$b \leftarrow 3 + (n/2) - \vec{d}$
Sample q from Beta(a, b)

Sample l_t (time scale) and σ_{tk}^2 by Metropolis-Hastings
Recreate B and d with NNGP rules, s.t. $\sigma_{tk}^{-2}\Sigma_{l_t}^{-1} + \sigma^{-2}I = B^\top \text{diag}(d)^{-1}B$
$V \leftarrow (B^\top \text{diag}(d)^{-1}B + \sigma^{-2}I)^{-1}$
Use Sparse Cholesky decomposition on V to compute and sample:
$m \leftarrow V(\sigma^{-2}(z - X\beta))$
$\vec{a} \leftarrow N(m, V)$

end for

Prediction Results



- During Hurricane Wilma, the model performs well at prediction, with the peak at both Marathon and Key West captured well within the predictive interval.
- However, since we assumed that the temporal noise is the same at all points, we see that the variance is overestimated during the more typical low surge times before and after the hurricane.

Model Interpretation

- Investigating the coefficients, we see that due to high collinearity between predictors at the current and adjacent hours, no coefficients appear to be significant. A few coefficients are shown below.

Coefficient	Mean	Median	95% CI	
Current Wind Speed	0.009	0.009	[-0.203187	0.210760]
Current Rain	-0.001	-0.001	[-0.053226	0.051910]
Current Pressure	0.011	0.011	[-0.376969	0.401043]
Previous Hour Wind Speed	-0.004	-0.004	[-0.135874	0.134129]
Previous Hour Rain	0.001	0.001	[-0.038250	0.036368]
Previous Hour Pressure	0.001	0.001	[-0.224543	0.210161]
Next Hour Wind Speed	0.004	0.004	[-0.128162	0.146859]
Next Hour Rain	-0.004	-0.004	[-0.038950	0.034374]
Next Hour Pressure	-0.017	-0.017	[-0.237471	0.204182]

- The RBF kernel time scale parameter sampled indicates that the temporally varying intercept is highly correlated with the previous 2-5 hours' intercepts.

GP Theory and Future Considerations

- Aad van der Vaart and Harry van Zanten have shown in 2011 that in response Gaussian Process models (where the response is directly modeled by a GP), under the RBF kernel, in both fixed and random designs, the posterior distribution of f_0, the true function of interest, converges to the true posterior as follows:

$$\mathbb{E}_{f_0}\left(\int ||f - f_0||^2 d\Pi(f|Y_{1:n})\right) \leq \frac{\ln(n)^{\frac{1}{r}}}{\sqrt{n}}$$

- Where f_0 restricts to $[0,1]^d$ a function with fourier transform $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ satisfying: $\int e^{\gamma||\lambda||^r}|\hat{f}|^2(\lambda)d\lambda < \infty$, for $r \geq 1, \gamma \geq 0$ (e.g. when $r = 1$, analytic functions on a strip of $\mathbb{C}^d$ containing $\mathbb{R}^d$)
- We are working on extending this result to the latent Gaussian Process (used here). We hope to achieve similar results by looking at the residuals from the other parts of the model, instead of investigating the response itself.
- It seems that to create more prescise predictive intervals, we may need to allow for the model variance to vary temporally.

References

- Aad van der Vaart and Harry van Zanten (2011). Information Rates of Nonparametric Gaussian Process Methods. Journal of Machine Learning Research vol 12, no. 60, 2095-2119.
- Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2016). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. Journal of the American Statistical Association, 111(514), 800–812.
- NOAA Tides and Currents (2024): Tides & Great Lakes Water Levels. National Oceanic and Atmospheric Administration.
- Florida Water Management Districts (2024): Hydrologic Conditions and Water Data. Florida Department of Environmental Protection.

hsin.huang@ucf.edu; john.scott2@ucf.edu