**A DATA-DRIVEN modeling framework for estimating BEACH sand grain size OVER regional scaleS**

D. Buscombe1, E. Goldstein2, K. Vos3, W. Speiser4, and S. Vitousek5

1Marda Science, contracted to U.S. Geological Survey, Pacific Coastal and Marine Science Center, Santa Cruz, CA, USA. Corresponding author email: [dbuscombe@contractor.usgs.gov](mailto:dbuscombe@contractor.usgs.gov)

2University of North Carolina at Greensboro, Greensboro, NC, USA.

3UNSW Water Research Laboratory, University of New South Wales, Sydney, Australia.

4University of California Davis Bodega Marine Laboratory, Bodega Bay, CA, USA.

5U.S. Geological Survey, Pacific Coastal and Marine Science Center, Santa Cruz, CA, USA.

1. **Introduction**

Grain size is a fundamental property of sand beaches, governing, for example, habitats, storm response and recovery, and morphodynamic state. There is a well-known relation between beach slope and grain size. Empirical models for beach grain size tend to either use beach slope as the only independent variable (e.g., Bujan et al., 2019), or both beach slope and wave climate statistics (e.g., McFall, 2019). These studies use linear regression on sparse datasets consisting of tens to hundreds of data points to represent all sand beaches globally. Here we describe an alternative approach to estimating regional trends in sand-beach grain-size using a Machine Learning regression technique and partial correlation analysis. As primary inputs, we use beach slope derived from satellite-derived-shoreline analyses, and wave statistics from 30km grid ERA5 wave monitoring (Hersbach et al., 2020), both of which may be obtained for any global location. Optionally, other regionally important covariates such as tide range and mineralogy may also be used. The approach also allows for quantitative ranking of grain-size covariates.

1. **Methods**

Empirical approaches to estimating the median sand grain size *D50* (mm), in a given location, traditionally derive from a linear least-squares model that is solved using a set of covariates *Xi*, such as slope (Bujan et al., 2019), slope and wave variables (McFall, 2019), or shoreline variability (Cabezas-Rabadán et al., 2021). Here we use a Gradient Boosted Regression Tree (GBRT) for the regression task of estimating *D50* from a suite of wave climate statistics obtainable from the ERA5 reanalysis such as peak wave period, *Tp*, mean wave direction, *θ*, and significant wave height, *H* (Hersbach et al., 2020*s*), mean spring tide range, *MSTR*, and satellite-derived shoreline trend, *S*, (Vos et al., 2019) and intertidal slope, *β*, (Vos et al., 2020). Our GBRT regressors are additive models. At each stage of gradient boosting, a decision tree (a so-called “imperfect estimator”), *hm*, is fit to the data. The model improves progressively by adding a new estimator, fitted to minimize a sum of losses (we use mean-square error), given the previous ensemble. We use a combination of feature engineering and partial correlation analysis to determine an optimal set of covariates, involving trials of select combinations of variables among candidate covariates. Covariates *Xi* are a) sampled using stratification (in the case study below, by latitude), b) standardized, and c) include a random number vector to assess the importance of variables relative to random.

1. **Case study: Southeast U.S.A. beaches**

To illustrate the modeling framework, existing grain-size data in the literature were compiled from 270 sand beaches sampled within the last 10 years, covering approximately 1800 km of shoreline along Florida, Georgia, South Carolina, North Carolina, and Virginia in southeast U.S.A. Figure 1a-f shows the results from 6 models trained on a reproducibly random draw of 25% of those data, in the form of observed versus estimated *D50* for the remaining 75% of data. The baseline models of Bujan et al., (2019) and McFall, (2019) perform less well (Figure 1g). The *β* and *MSTR* model (Figure 1c) has a R2 of ~0.8 and a mean error of ~12%. A repeated *K*-fold cross-validation (*K* = 10, number of repeats = 5) of that model yields a mean squared error of 0.04 mm. A less-parsimonious model constructed using *β, MSTR, Hs, Tp,* and *θ* (Figure 1f) has a higher error based on the test data. In this region, *MSTR* is the most important variable for prediction.

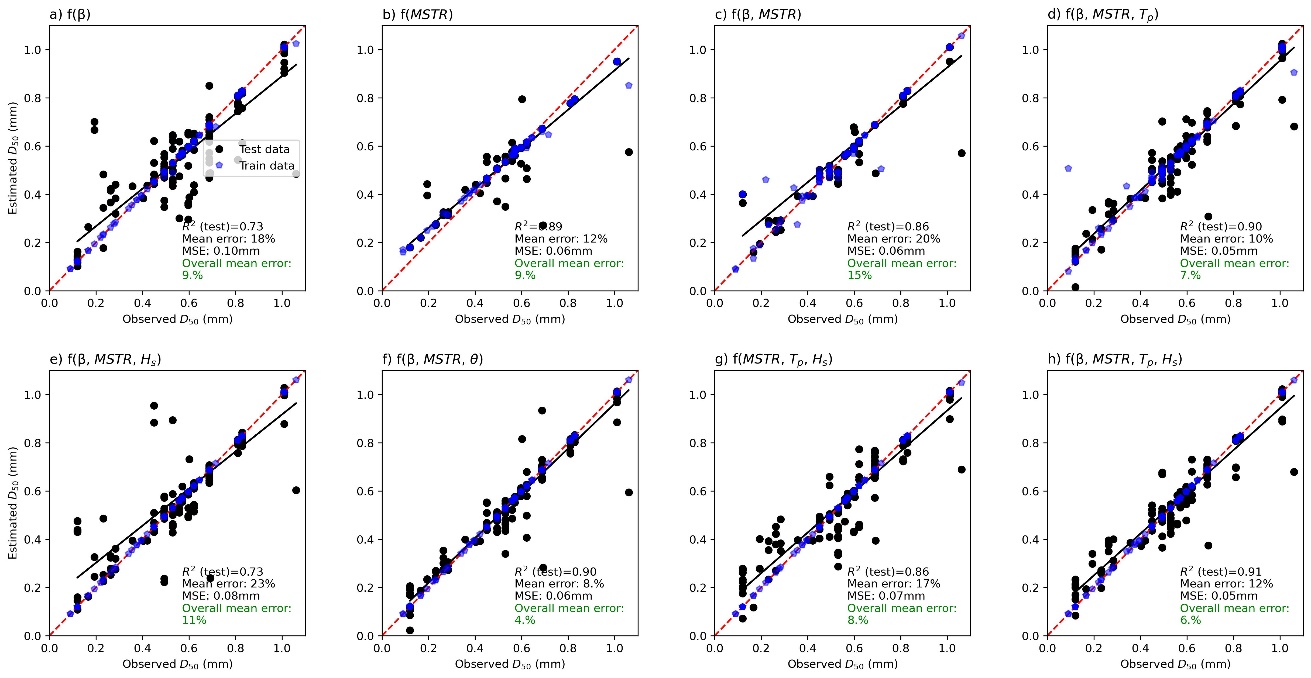


Figure 1. Observed versus estimated *D50* in the Southeast USA, using a variety of models (panels a through h) that differ in terms of model inputs *Xi*

1. **Discussion**

Grain size on sand beaches is well-known to vary at event scales (e.g. Masselink et al., 2008), therefore observational data are always preferable for management and scientific objectives. The model framework described here might find useful applications in lieu of or as a supplement to such field data, and to inform and stimulate the design of further grain size sampling. We are currently collating data to apply the modeling framework to other regions to test its general utility. We expect that each model will be unique to each region, owing to differences in relative wave and tidal regimes, sediment supply, sediment mineralogy, and carbonate content. While potentially valuable in regions in which such models may be applied, this work may also benefit the community if it leads to a greater fundamental understanding of what controls beach grain size, and if it can stimulate the community to sample more, especially where model predictions are inaccurate. Models should only improve with more, and better, grain size data.

**References**

Bujan, N., Cox, R., and Masselink, G., 2019. From fine sand to boulders: Examining the relationship between beach-face slope and sediment size, *Marine* *Geology*, 417, 106 012

Cabezas-Rabadán, C., Pardo-Pascual, J. E., and Palomar-Vázquez, J., 2021. Characterizing the Relationship between the Sediment Grain Size and the Shoreline Variability Defined from Sentinel-2 Derived Shorelines, *Remote Sensing*, 13, 2829.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz‐Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D. and Simmons, A., 2020. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), pp.1999-2049.

Masselink, G., Buscombe, D., Austin, M., O'Hare, T. and Russell, P., 2008. Sediment trend models fail to reproduce small‐scale sediment transport patterns on an intertidal beach. *Sedimentology*, 55(3), pp.667-687.

McFall, B. C., 2019. The relationship between beach grain size and intertidal beach face slope, *Journal of Coastal Research*, 35, 1080–1086.

Vos, K., Splinter, K. D., Harley, M. D., Simmons, J. A., and Turner, I. L., 2019. CoastSat: A Google Earth Engine-enabled Python toolkit to extract shorelines from publicly available satellite imagery, *Environmental Modelling & Software*, 122, 104 528,

Vos, K., Harley, M. D., Splinter, K. D., Walker, A., and Turner, I. L., 2020. Beach Slopes From Satellite-Derived Shorelines, *Geophysical Research Letters*, 47, e2020GL088 365