

1 **A Review of Machine Learning Applications to Coastal Sediment Transport and**  
2 **Morphodynamics**

3

4 Evan B. Goldstein\*<sup>1</sup>, Giovanni Coco<sup>2</sup>, Nathaniel G. Plant<sup>3</sup>

5 <sup>1</sup>Department of Geological Sciences, University of North Carolina at Chapel Hill, 104 South Rd,  
6 Mitchell Hall, Chapel Hill, NC 27599 USA

7 <sup>2</sup>School of Environment, University of Auckland, New Zealand

8 <sup>3</sup> U.S. Geological Survey St. Petersburg Coastal and Marine Science Center, St. Petersburg,  
9 Florida, USA

10 \*Corresponding Author: [evan.goldstein@unc.edu](mailto:evan.goldstein@unc.edu) (email)

11 [@ebgoldstein](https://twitter.com/ebgoldstein) (Twitter)

12

13 **Abstract**

14 A range of computer science methods under the heading of machine learning (ML) enables  
15 the extraction of insight and quantitative relationships from multidimensional datasets. Here, we  
16 review some common ML methods and their application to studies of coastal morphodynamics  
17 and sediment transport. We examine aspects of ‘what’ and ‘why’ ML methods contribute, such  
18 as ‘what’ science problems ML tools have been used to address, ‘what’ was learned when using  
19 ML, and ‘why’ authors used ML methods. We find a variety of research questions have been  
20 addressed, ranging from small-scale predictions of sediment transport to larger-scale sand bar  
21 morphodynamics and coastal overwash on a developed island. We find various reasons justify  
22 the use of ML, including maximize predictability, emulation of model components, smooth and  
23 continuous nonlinear regression through data, and explicit inclusion of uncertainty. Overall the  
24 expanding use of ML has allowed for an expanding set of questions to be addressed. After  
25 reviewing the studies we outline a set of ‘best practices’ for coastal researchers using machine  
26 learning methods. Finally we suggest possible areas for future research, including the use of  
27 novel machine learning techniques and exploring ‘open data’ that is becoming increasingly  
28 available.

29

30 **1. Introduction**

31 The amount of available data on coastal systems has increased dramatically in recent years,

32 ranging from topographic and bathymetric data (e.g., Turner et al., 2016), to compilations and  
33 collections of sediment transport and physical forcing (e.g., Bolaños and Souza, 2010; Garel and  
34 Ferreira, 2015; Nelson et al., 2013; van der Werf et al., 2009). Large spatial and temporal  
35 extents, high resolution, and rapid turnaround from acquisition to availability means that the data  
36 being produced enables expanded applications to coastal morphodynamic research. In particular,  
37 since observational data has always been the foundation for developing empirical relationships or  
38 testing quantitative models, the recent volume of data available, the intrinsic high dimensionality  
39 and nonlinearity of underlying processes, and increased computing power, have all led to  
40 renewed interest in empirical research.

41 A key example in this new wave of research is attempting to extract insights, predictions,  
42 or quantitative relationships directly from multidimensional datasets using automated tools. This  
43 ‘data-driven’ route for science has been demonstrated to be a promising research direction (e.g.,  
44 Anderson, 2008; Hey et al., 2008), and tools from a range of disciplines have been influential in  
45 defining and tackling ‘data-driven’ science (e.g., Staltzer and Mentzel, 2016). In this review we  
46 differentiate classic empirical work and this new wave of ‘data-driven’ work as being divided by  
47 the computational methodology, as well as (potentially) the quantity, nonlinearity expressed in  
48 the data, and high dimensionality. Our focus is on the new empirical work using machine  
49 learning, the set of computer algorithms, methods and tools that implement a given task and use  
50 data to optimize performance (e.g., reduction of error). In this manuscript we provide many  
51 examples of successes in the use of machine learning for coastal research, but first we discuss the  
52 rationale for this data-driven approach.

53 Data-driven research is inductive. As with other empirical work, data-driven research relies  
54 on data to develop insight, predictions, or relationships. We acknowledge that empirical work  
55 does not and cannot exist in a vacuum — theory and logic are critical parts of data analysis (e.g.,  
56 Coveney et al., 2017; Crutchfield 2014) and mathematical proofs show the lack of  
57 generalizability of inductive statements (e.g., Popper and Miller, 1983). However, inductive  
58 statements are part of the scientific workflow, and are unavoidable at certain junctures. Even  
59 Newton expressed the utility of induction in ‘Rule 4’ in the 3rd edition of the Principia (Cohen et  
60 al., 2016):

61

62 *“In experimental philosophy, propositions gathered from phenomena by induction should be*  
63 *considered either exactly or very nearly true notwithstanding any contrary hypotheses, until yet*  
64 *other phenomena make such propositions either more exact or liable to exceptions.”*

65

66 Coastal morphodynamics specifically and geomorphology in general have long histories  
67 of induction, and developing empirical rules that are useful. Even when basic laws of physics can  
68 be used, empirical expressions or rules of some form or another are always required to close an  
69 equation set. For example, sediment transport rules, turbulent closure schemes, friction  
70 coefficients, and wave breaking all rely on ad hoc rules, assumptions or empirical relations. If we  
71 still need inductive rules, how should we build them? With the increased quantity of data, and  
72 the improvements in computing power, coastal researchers have access to a wide range of  
73 computer science tools from the subdiscipline of machine learning (ML) to develop inductive  
74 statements and optimized predictions directly from data sets.

75 Much of the machine learning work we discuss in this review is inherently focused on  
76 identifying and exploiting correlations and patterns in data. Assigning causation can be less clear  
77 in some coastal morphodynamic systems because of multiple scales (ripples, megadunes, bars,  
78 shoreline), and feedbacks between scales (bars impact shoreline, and vice versa) that interact in  
79 both space and time (e.g., Murray et al., 2014a, 2014b; Sherman, 1995; Short et al., 1985;  
80 Werner, 1999; Winant et al., 1975). However, correlation is valuable for prediction because of  
81 the concept of analogy (Lorenz, 1969a; 1969b) — knowing how a coastal system evolved when  
82 it was in the same configuration but at a previous time can lead to predictions about how the  
83 system might evolve in the future (i.e., seasonal dynamics of sediment transport; Aubrey, 1979;  
84 Plant et al., 1999; Plant et al., 2006; Splinter et al. 2011; Yates et al., 2009). A data-driven  
85 approach can help to elucidate this behavior by examining previously collected data, and  
86 developing a model focused on how the system will evolve based on past instances.

87 Because empirical approaches rely on the data to make predictions about a system, data-  
88 driven work may only be strictly applicable within the range of the data used to develop the  
89 predictor — unless the prediction scheme can be argued to be more generally valid. This is a  
90 limit of all inductive, empirical techniques, though it is rarely mentioned in more traditional  
91 empirical studies (i.e., any study that uses linear regression to predict beyond the bounds of the  
92 data). Furthermore, this caveat is likely applicable to all modeling studies because new processes

93 or feedbacks can exist that are not included within the model. A morphodynamicist might argue  
94 that, if a given model is built from conservation laws, the model should be able to predict outside  
95 of the range of conditions where the model has been tested. This argument also holds for data-  
96 driven work — the data used to construct the model adheres to conservation laws, therefore  
97 predictions (built directly from data) might also adhere to these physical constraints outside of  
98 the range of data used to build the model. Our point here is to suggest to readers that data-driven  
99 work should not be disregarded because it is not built from ‘laws’ — data-driven work is based  
100 on data, which obey conservation laws.

101 While there are a number of aspects of coastal science that can and do benefit from  
102 machine learning, we focus here on predictions of coastal sediment transport, coastal  
103 morphology, and coastal morphodynamics. The review here is focused on supervised learning,  
104 specifically regression tasks using continuous data (as opposed to classification tasks).  
105 Supervised learning involves input and output data that are linked (such as wave forcing and a  
106 given morphological configuration) with the goal of developing a function to relate the input to a  
107 corresponding output and emulating physical processes relationships that are either poorly  
108 understood or complex and difficult to capture with deterministic models. Excluded here are  
109 prediction of forcing and fluid phenomena when no reference to sediment transport is given or  
110 studies focused on engineering and structures.

111 Previous ML work written for an Earth and Environmental Science audience has focused  
112 primarily on introducing ML algorithms, in the form of both books (Hsieh, 2006) and papers  
113 (Chau, 2006; Valentine and Kalnins, 2016). This previous work serves a key role in connecting  
114 coastal scientists to ML tools (e.g., Jones and Maccarone, 2013). Our work intends to move the  
115 purview of these previous reviews, which focused on introducing ML algorithms, explaining ML  
116 algorithms, and providing a few select examples to review. In this document, we do provide a  
117 brief introduction to the most common machine learning techniques that have been used in  
118 coastal sediment transport and the steps needed to start a ML project (Section 2). However, our  
119 focus is on comprehensively reviewing previous machine learning work on coastal  
120 morphodynamics such that these works can be recognized, compared, and used to build future  
121 ML efforts. To this end we review and discuss more than 60 papers under three separate  
122 headings: studies where ML is used to predict sediment transport (Section 3); studies where ML  
123 is used to make a stand-alone morphodynamic model (Section 4); and studies where ML is

124 embedded or linked to a morphodynamic model (Section 5). Furthermore, we address ‘why’ ML  
125 tools have been used in particular studies and what was learned by using ML methods. We  
126 intend coastal morphodynamicists interested in ML to use these sections (3, 4, and 5) to assess  
127 what research has been done (and what has not been done) at the time of this writing. This  
128 comprehensive review of the coastal ML literature also permits us (in Section 6) to discuss  
129 overarching topics, offer a set of best practices for open, reproducible, replicable machine  
130 learning research and highlighting some future directions in coastal ML research primarily  
131 focused on synthesis and intercomparison.

132

## 133 **2. Machine learning methods used in Coastal sediment transport and Morphodynamics.**

134 Before we review the uses of ML in coastal science, we introduce the ML methods such  
135 that sections (Sections 2.1 - 2.6) provide basic information on each ML method. Within each of  
136 these sections we provide relevant papers for readers who wish for more details regarding each  
137 method

138

### 139 **2.1 Artificial Neural Network (ANN)**

140 Artificial Neural Networks (ANN) are commonly used algorithms in machine learning  
141 because of their versatility. Many different fields of science have used ANNs for tasks such as  
142 function fitting to classification. Applications in coastal sediment transport and  
143 morphodynamics include multiple aspects of suspended sediment transport, sandbar  
144 morphodynamics, and various studies of shoreline position — all mentioned in the following  
145 sections. The most typical form of an ANN is represented by a series of layers: an input layer,  
146 one or more hidden layers and an output layer. Each layer consists of a number of nodes  
147 (artificial neurons). The input data is fed to the network via a node on the input layer (usually  
148 each node represents an input variable) while, depending on the number of variables to be  
149 predicted, the output layer could consist of one or more nodes. The hidden layer(s) contain a  
150 somewhat arbitrary number of nodes chosen based on a mix between experience, empirical  
151 formulas and systematic analysis. An idealized feed-forward ANN characterized by  $n$  input  
152 nodes (the predictors or independent variables),  $m$  hidden nodes and one output node (the  
153 prediction or dependent variable). Nodes are mathematically connected and transfer information  
154 from the input variables to a node of the hidden layer:

155

$$156 \quad h_j = f \cdot \left( a_j + \sum_{i=1}^n w_i x_i \right) \quad (1)$$

157

158 where  $x_i$  is the  $i^{\text{th}}$  of  $n$  input variables,  $h_j$  the response of the  $j^{\text{th}}$  neuron in the hidden layer,  $f$  is the  
159 activation function (e.g., a sigmoid, an hyperbolic tangent, etc.),  $w_i$  is the connection weight  
160 between  $x_i$  and  $h_j$ , and  $a_j$  is the bias for the  $j^{\text{th}}$  hidden neuron. A further combination of the hidden  
161 nodes, which is achieved by means of a new activation function (not necessarily the same as the  
162 one used to link the input variables and the hidden layer) and new connection weights and biases,  
163 connects the hidden layer to the output layer.

164 The biases and connection weights of the ANN are established through an optimization  
165 algorithm that is applied to a dataset consisting of observed input and output variables. Various  
166 algorithms can be used to perform this critical step, though it remains difficult to tell *a priori*  
167 which optimization function will provide better results. Many of these algorithms are based on  
168 the backpropagation of errors — the error at the output (prediction) layer is sent back through the  
169 network to adjust and update the weights and biases.

170 ANNs are often portrayed as an example of a black-box predictor where the (usually) large  
171 number of weights and biases obscures the role of individual variables. The architecture of  
172 small-size ANNs can in fact be analyzed and various techniques have been developed to this aim  
173 (e.g., Olden et al., 2004). LeCun et al. (2015) provide many helpful references and a relatively  
174 recent review on ANN that focuses on current research themes (i.e., ‘Deep Learning’; ANNs  
175 with many hidden layers).

176

## 177 **2.2 Genetic Algorithms (GA) and Genetic Programming (GP)**

178 Genetic algorithms (GA; Holland, 1975) and genetic programming (GP; Koza 1992) are  
179 related ML techniques that operate on rules based on natural selection. In the section below we  
180 review the basics of genetic algorithms. For example, consider an equation with five free  
181 parameters, where a given combination of specific values for each parameter can be compressed  
182 into a single string of length 5. The string of parameter values is also related to the solution of the  
183 equation using these 5 specific values. Each solution is also related to an associated error (the  
184 value of the equation using the 5 parameters vs. some measured value). Now consider a

185 population of such strings (not just one), each with their own unique combination of values for  
186 each of the 5 parameters. The genetic algorithm routine works by operating on these strings  
187 using evolutionary rules. Given an initial set of strings (a population), there is an error associated  
188 for each string. The strings with the smallest error are retained; the strings with the most error are  
189 discarded. New strings are developed by mutation (changing values in a given string) and by  
190 reproducing — recombining two strings to make a novel new string. By using these  
191 ‘evolutionary’ rules, the routine will search over the solution space and tend to converge on  
192 solutions that are globally optimal. Parameters in the evolutionary rules and techniques in  
193 applying those rules are tunable (e.g., number of predictors in the population, mutation percent,  
194 crossover rules, number of generations, number of discarded or kept predictors for each  
195 generation, etc.). Genetic algorithms can be used in tandem with artificial neural networks — for  
196 example to find the appropriate weights and biases, as well as network architecture (e.g., Yao,  
197 1999). Further helpful entries into the GA literature can be found in D’Ambrosio et al. (2013),  
198 Mitchell (1995) and Mitchell (1998).

199 Building on the population approach of genetic algorithms, genetic programming (GP;  
200 Koza, 1992) takes the idea a step further. The population is not strings of parameter values to be  
201 input into a fixed equation but instead a population of equations with mutable form and length.  
202 Given a set of mathematical operators (+, -, \*, /), and a set of input variables (e.g., forcing  
203 conditions) a GP routine works to find equations using these building blocks (input variables,  
204 constants, and mathematical functions) — this is a symbolic regression problem. One issue with  
205 GP is the development of large, complex functions that have small error compared to small, less  
206 complex functions that have larger error but might be more physically interpretable. Therefore  
207 routines may offer more than one solution, and instead offer many solutions to the problem  
208 which fall along the Pareto front — a line in error-complexity space that defines how prediction  
209 error decreases with the solution complexity (a measure of the size of the predictor that  
210 incorporates the mathematical operators, variables, and constants). The act of choosing a  
211 predictor from this front introduces subjectivity in the routine, though GP algorithms have shown  
212 the ability to find physically meaningful results from data streams. Aside from the work of Koza  
213 (1992) introducing the technique, the book by Poli et al. (2008), and work by Babovic and  
214 Keijzer (2000), Olden et al. (2008), Schmidt and Lipson (2009), and O’Neill et al. (2010) have  
215 proven helpful to us.

216

### 217 **2.3 Bayesian Networks (BN)**

218 Bayesian networks (BN) implement a form of probabilistic prediction that explicitly  
219 resolves the conditional probabilities that link variables to one another, albeit in a discretized  
220 fashion. Statistical operations include marginalization over a subset of a larger distribution, for  
221 instance, when the data are used to provide constraints (Charniak, 1991). And, as the name  
222 suggests, Bayesian estimation (Cooper and Herskovits, 1992; Malakoff, 1999) can be  
223 implemented to solve problems that typically require data assimilation (Wikle and Berliner,  
224 2007). For example, to estimate coastal erosion that is assumed to be influenced by dune  
225 morphology, geology, and sea-level rise (Plant et al., 2016) the probabilistic relationship can be  
226 expressed as:

$$227 \quad P(E_i) = \sum_{G,D,SLR} P(E_i | D, G, SLR) P(D | G, SLR) P(G, | SLR) P(D) P(SLR) \quad , (2)$$

228 where left side of the equation is describing the probability that a certain amount of erosion,  $E_i$ , is  
229 experienced. The right side of the equation is the product of the conditional probability of that  
230 amount of erosion occurring, given the morphologic state of coastal dunes ( $D$ ) and geologic  
231 setting ( $G$ ) and a sea-level rise rate ( $SLR$ ). The probability is integrated over all the states, which  
232 may be constrained by data. This is the marginalization operation. Some of the terms on the  
233 right side of equation (2) defining the erosion probability may themselves have dependencies  
234 that can be solved using Bayes rule:

$$235 \quad P(D_i) = \sum_E P(E | D_i, G, SLR) P(D_i) / P(E), \quad (3)$$

236 where the first term on the right side of the equation (2) is inverted. Bayes rule and  
237 marginalization can take place simultaneously in a Bayesian network, implying that there is no  
238 real distinction between a forward implementation that emulates a deterministic model (e.g., a  
239 partial differential equation) and an inverse model.

240 The approach models probabilities directly, as opposed to modeling the process-variables  
241 as is done in the other ML examples. This is useful if knowing the uncertainties is a primary  
242 modeling requirement. A disadvantage is that the model must learn the conditional probabilities  
243 that describe the correlations between variables, and this comes with a cost of increasing free  
244 parameters that grows as the number of states raised to the number of variables. Furthermore the  
245 uncertainty present in the resulting model only reflects the uncertainty that is found within the  
246 data. We have found general papers by Aguilera et al. (2011), Chen and Pollino (2012), and



247 Uusitalo (2007) to be useful in learning techniques and applications of BN.

248

## 249 **2.4 Regression Trees (RT)**

250 Regression trees (RT) separate prediction tasks into a series of binary splits, leading to a  
251 branching, tree-like structure (e.g., De'ath and Fabricius, 2000; Hastie et al. 2009). One  
252 advantage of these tree-based approaches is that they easily allow users to assess the relative  
253 influence of the input variables. Trees are visually appealing and reading through a RT model  
254 can be straightforward, especially when the tree is short.

255 An example of a regression tree-based algorithm is recursively splitting the dataset into  
256 groups. The details of each split are determined via a given metric, such as minimizing the sum  
257 of squares of each group. A variety of rules exist for both growing trees (i.e., how many  
258 recursive splits) and pruning trees (removing splits). Additionally, other algorithms can be  
259 attached to tree based methods to improve accuracy, specifically 'Boosting'. Boosting routines  
260 merge many small regression tree models that are built sequentially, with misfit data sequentially  
261 given more weights so trees progressively focus on poorly predicted data (Elith et al., 2008). We  
262 have found the works of De'ath and Fabricius (2000), De'ath (2007), Olden et al. (2008), and  
263 Hastie et al., (2001), useful for learning about these tree based approaches.

264

## 265 **2.5 Nonlinear forecasting (NF)**

266 Though it may not strictly be classified as a ML method, nonlinear forecasting (NF) has  
267 an affinity to machine learning methods, so we discuss it in this context. Nonlinear forecasting is  
268 built from autoregressive models for predicting time series:

269

$$270 S_t = a_0 + \sum_{i=1}^m a_i S_{t-(i\Delta t)} \quad (4)$$

271

272 where  $a_0$  through  $a_m$  are coefficients,  $S$  is the variable of interest,  $t$  is time, and  $m$  is the number of  
273 past time instances (with temporal spacing of  $\Delta t$ ) used to develop an evaluation of  $S$  at time  $t$ . If  
274 the coefficients  $a_0$  through  $a_m$  are constant for the entire time series, the model is global and  
275 linear. If the coefficients vary as a function of  $S$ , then the model is nonlinear, (i.e., only locally  
276 linear). Coefficients and their variation as a function of  $S$  can be determined through nearest

277 neighbor approaches. There are additional ways to introduce nonlinearity, such as deciding a  
278 specific nonlinear function for the right hand side of (4), or incorporating a threshold.

279 In situations where there is a dependence on initial conditions and external forcing, the  
280 system can be modeled using approaches such as input-output models (Casdagli, 1991). Adding  
281 external forcing inputs to the autoregressive framework is the basis of ARX models  
282 (AutoRegressive model with eXogenous inputs):

283

$$284 \quad S_t = a_0 + \sum_{i=1}^m a_i S_{t-(i\Delta t)} + \sum_{k=1}^n b_k Q_{t+(k\Delta t)} \quad (5)$$

285

286 where  $Q$  represents an external forcing signal,  $b_1$  through  $b_n$  are coefficients, and  $k$  is the number  
287 of past instances (with temporal spacing  $\Delta t$ ) used to develop the model. Coefficients again can  
288 vary as to make the model nonlinear (via locally linear sections). Other modifications to this  
289 framework include NARX models (Nonlinear AutoRegressive models with eXogenous inputs),  
290 which adjust the functional form of the right hand side of the multiple linear regression equation  
291 (5). For example, a tuned neural network can be used as the right hand side of (5), an example of  
292 a NARX model.

293 Key recent papers to understand the technique and application of nonlinear time series  
294 analysis, nonlinear forecasting, and other related work (under the heading of ‘Empirical Dynamic  
295 Modeling’) are Casdagli (1989), Chang et al. (2017), Farmer and Sidorowich (1987), Kantz and  
296 Schreiber (2004), Packard et al. (1980), Sugihara et al. (2012), Sugihara and May (1990), Ye and  
297 Sugihara (2016), and finally Takens (1981), who provided much of the theoretical background  
298 on which the techniques are based.

299

### 300 **3. Applications to Coastal Sediment transport**

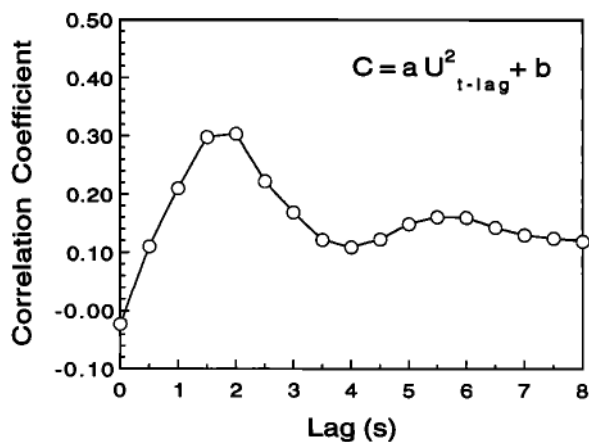
301 We have now reviewed the most used ML techniques in coastal morphodynamics and  
302 sediment transport studies. The availability of coastal sediment transport data, and the lack of a  
303 single ‘perfect’ predictor (for a given sediment transport relation) has led to the hope that ML  
304 will provide a more viable, optimal sediment transport equation — a motivation for many of the  
305 works that we review in this section. Authors frequently want to develop a predictor that is either  
306 more generally valid (better prediction with a large set of data) or more specifically valid (better  
307 prediction with a small set of data specifically collected for a given setting/condition). Authors of

308 the studies reviewed below all test their ML prediction scheme against established predictors  
309 from the literature (i.e., previous empirical or theoretical sediment transport prediction schemes).  
310 The newly developed ML techniques often performs better than the traditional scheme using the  
311 error metric selected by the authors, a phenomena we discuss in Section 6.1.

### 312 3.1 Suspended Sediment Concentration

313 Predictions of suspended sediment concentration are a fundamental test of theoretical and  
314 statistical understanding of sediment mobility and transport that control morphologic evolution  
315 on a wide range of spatial and temporal scales. Time-varying sediment concentrations have been  
316 predicted using several ML methods. Jaffe and Rubin (1996) used nonlinear forecasting  
317 techniques to predict suspended sediment concentration based on instantaneous water velocity  
318 (with and without higher order velocity terms) and various water velocity history terms (e.g.,  
319 velocity at the previous time step, etc.). A notable aspect of this study is the investigation of the  
320 appropriate time lag in water velocity to maximize the correlation with sediment transport  
321 (Figure 1) an insight that may have transferability to other studies on time lags in coastal  
322 systems.

323



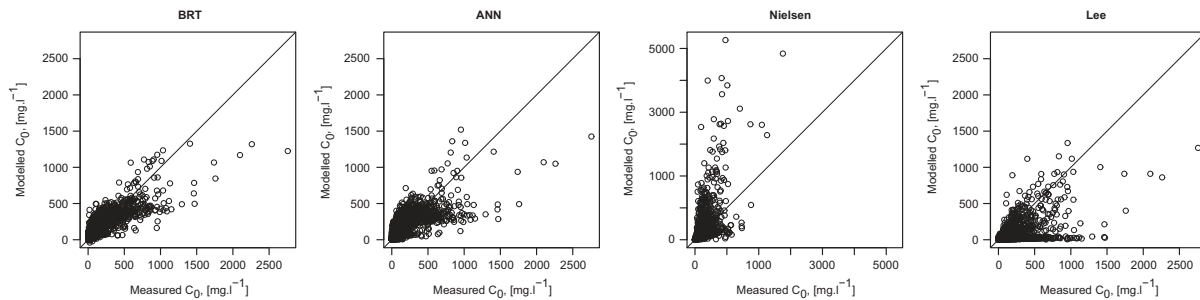
324

325 *Figure 1: A plot from Jaffe and Rubin (1996), who used nonlinear forecasting to predict*  
326 *suspended sediment concentration under waves. The plot above exhibits the changing*  
327 *correlation between suspended sediment and wave forcing with changing lag time on the wave*  
328 *forcing term.*

329

330 As an extension to this work, Yoon et al. (2013) used an ANN to predict time-dependent

331 suspended sediment concentration as a function of various hydrodynamic parameters both inside  
332 and outside the surf zone. With such a large dataset and many measured variables, Yoon et al.  
333 (2013) was able to use the ANN to identify the hydrodynamic parameters (and combinations of  
334 parameters) that are most predictive in different regions of the laboratory surfzone. Using a GP  
335 routine, Kizhisseri et al. (2005) used both synthetic and field data to produce expressions for  
336 suspended sediment concentration based on instantaneous fluid velocity (and higher powers of  
337 velocity). In a rare example of a reported unsuccessful ML application in coastal  
338 morphodynamics, the prediction of suspended sediment concentration using field data lead to  
339 poor performance (i.e., a large absolute error for the prediction; Kizhisseri et al., 2005). Oehler et  
340 al. (2011) used both BRT and ANN to develop predictors for near bed suspended sediment  
341 reference concentration based on water depth, median grain size, mean wave period at the bed,  
342 wave orbital amplitude at the bed, and significant wave orbital speed at the bed. The BRT model  
343 was superior to ANN (Figure 2), which we highlight because many studies do not compare ML  
344 derived predictors developed from multiple ML routines. Oehler et al. (2011) provides a clear  
345 example that this work should be done, and will inform future researchers wondering about  
346 which ML method to use to predict suspended sediment reference concentration. Goldstein et  
347 al., (2014) used the same dataset and developed a GP routine to construct a predictor for  
348 reference concentration. This predictor was specifically derived for use in a numerical model of  
349 inner shelf bedforms (discussed further in Section 5), and is an example of a predictor developed  
350 to work in a specific (multiple grain size) setting. At a larger scales, Teodoro et al. (2007) used  
351 an ANN on remotely sensed images of the surf zone to predict total suspended matter in the  
352 water column along the coast of Portugal based on satellite remote sensing data (calibrated with  
353 field measurements of seawater reflectance). This work highlights the potential for predicting  
354 suspended sediment concentration using the high temporal and spatial resolution remote sensing  
355 products available, which could potentially be linked to global measures of shoreline change  
356 (e.g., Luijendijk et al., 2018).



357

358 *Figure 2: A figure from Oehler et al. (2011) exhibiting the performance of a BRT and ANN*  
359 *model for suspended sediment reference concentration compared to two more traditional*  
360 *prediction schemes.*

361

### 362 **3.2 Suspended Sediment Flux**

363 Scaling up from instantaneous concentration to alongshore-directed suspended sediment  
364 flux has been the focus of several studies. Using an ANN, van Maanen et al. (2010) predicted the  
365 depth integrated alongshore sediment transport using water depth, wave height, wave period and  
366 alongshore current velocity. Analyzing the parameterized ANN also allowed van Maanen et al.  
367 (2010) to understand which parameters held the most explanatory power (alongshore current of  
368 velocity), and to understand when the predictor provided unphysical answers. Notably,  
369 unphysical predictions were found when the ANN was given input parameters outside the range  
370 of the training data, highlighting the importance of training models with extreme conditions.  
371 Predictors for the net alongshore sediment transport rate based on wave height, wave period,  
372 breaking wave angle, beach slope, and grain size have been developed using ANN (Kabiri-  
373 Samani et al. 2011) and regression trees (Mafi et al. 2013). The ability for both ML methods to  
374 produce successful predictors highlights the need for more comparative work between ML  
375 methods.

### 376 **3.3 Sediment Properties**

377 Finally there have been ML studies of sediment properties (e.g., mean grain size, skewness,  
378 kurtosis, fall velocity, etc.). Nysten et al. (2015) trained a decision tree to determine several  
379 aspects of beach and dune sediment in Finland as a function of environmental variables (e.g.,  
380 elevation, slope, curvature, local fetch, geography and climate conditions). Sediment parent  
381 material (parameterized via geography) was found to be an important control on grain size and  
382 sediment sorting, obscuring the role of local controls. Goldstein and Coco (2014) used a GP

383 routine to develop a predictor for noncohesive sediment settling velocity that incorporates fluid  
384 kinematic viscosity, relative sediment density and sediment nominal diameter. The study focused  
385 on the role of training dataset size and selection method while developing a prediction scheme  
386 that performed better than two common equations.

387

#### 388 **4. ML Morphological and Morphodynamic Models**

389 A variety of coastal morphology and morphodynamic models have been built using ML.  
390 Many researchers use ML as an optimization tool — looking for better morphological prediction  
391 with newly collected or existing data.

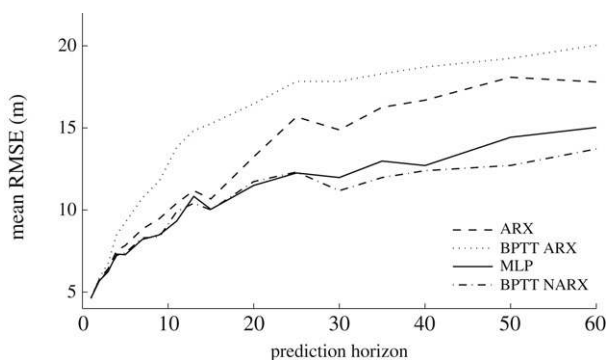
##### 392 **4.1 Sandbars**

393 Sandbar morphology (e.g., the cross-shore position and alongshore uniformity) has been a  
394 common focus of machine learning studies. Múnera et al. (2014) developed an ANN to  
395 determine the correlation between sandbar morphology and a given wave climate, culminating in  
396 in examining the nonlinear dependencies of bar position on past wave conditions (i.e., time-  
397 lagged wave conditions). Lopez et al. (2017) used an ANN to determine the cross-shore bar  
398 position given wave characteristics, sediment characteristics, and temporal data (month and day  
399 information). Compared to a common formula to predict bar characteristics, the optimized ANN  
400 had lower error. Komurcu et al. (2013) used an ANN to predict the geometric and shape  
401 characteristics of experimentally simulated bars based on the wave height, wave period, bed  
402 slope, and grain size. Tests were performed varying the split of training data/testing data. The  
403 best fit model was trained with the largest amount of data, and had lower error compared to  
404 literature formula. A similar study on experimental bar data was performed by Demirci et al.  
405 (2015), using wave parameters, bed slope, and sediment characteristics to predict bar volume  
406 using an ANN and multiple linear regression. The predictor derived from ANN outperformed the  
407 multiple linear regression.

408 Of particular note is the work of Pape et al., (2007; 2010), who used a recurrent artificial  
409 neural network to model the cross-shore position and temporal dynamics of sandbar crests.  
410 Recurrent neural networks are ANNs that feed output predictions back to the input layer of the  
411 ANN, making a forward in time morphodynamic model. Pape et al. (2007) modeled sandbar  
412 position using relevant wave inputs and previous sandbar positions using a linear autoregressive  
413 model with exogenous inputs and a recurrent neural network (i.e., a nonlinear autoregressive

414 model with exogenous inputs; NARX) trained using multiple techniques. All models exhibit  
415 decaying performance as the prediction horizon (the prediction lead time) increases, but  
416 nonlinear ANN models show slightly better results over long prediction timescales (Figure 3).  
417 Assessment of prediction timescale is especially critical to understand if data-driven techniques  
418 can be successful techniques for forecasting future morphology and morphodynamics, and  
419 understanding how error compounds or decays through time in data-driven models. Additionally,  
420 comparative work between data-driven methods is particularly interesting to the user of ML, and  
421 can give insight into inherent predictability and nonlinearity in the study system. Pape et al.  
422 (2010) continued this work, using two neural networks to model sandbar behavior and compared  
423 results to a traditional cross-shore morphodynamic model. Both data-driven models showed  
424 increased performance compared to the morphodynamic model, measured using the metric of  
425 error over increasing prediction horizon.

426



427

428 *Figure 3: The increase in sandbar position error for increasing prediction timescales using*  
429 *data-driven models (from Pape et al., 2007). Error is minimized when using models based on*  
430 *ANN (the ‘MLP’ and ‘BPTT NARX’ models).*

431

#### 432 **4.2 Shoreline position and shore profile**

433 Various shoreline attributes have also been predicted using machine learning techniques.  
434 Using an ANN to predict beach profiles (from the dunes to MSL) with wind and wave data,  
435 Hashemi et al. (2010) also discusses the role of training data that spans a wide range of conditions  
436 to avoid error associated with out-of-sample prediction. Grimes et al. (2015) analyzed beachface  
437 and shoreline timeseries data, using a GP and nonlinear forecasting to predict the dynamics of

438 intertidal beachface geometry and examine the role of internal dynamics vs. external controls  
439 (i.e., forcing). Both Tsekouras et al. (2015) and Rigos et al. (2016b) described new methods to  
440 formulate and train a novel ANN architecture to predict shoreline characteristics — Tsekouras et  
441 al. (2015) examined shoreline erosion as a function of storm characteristics and bathymetry,  
442 while Rigos et al. (2016b) investigated multiple shoreline positions and shoreline rotation given  
443 hydrodynamic inputs and offshore reef morphology. Iglesias et al. (2009a; 2010) used an ANN  
444 to predict the planform morphology of headland-bay-beach systems (including those with shore  
445 protection structures). Iglesias et al. (2009a) tested multiple ANN architectures (number of  
446 hidden layers and nodes) as well as different algorithms to train the ANN. The final ANN model  
447 outperformed previously developed shoreline models, and error from the ANN model was  
448 distributed across the shoreline as opposed to the previous models, which had concentrated zones  
449 of high error. Iglesias et al. (2009b) extended this work by incorporating tidal range into the  
450 ANN. After testing various ANN architectures and finding the best ANN predictor, Iglesias et al  
451 (2009b) used the trained model to examine the interplay between tidal range and wave  
452 parameters in controlling headland bay geometry. Loureiro et al. (2013) used a BN to  
453 probabilistically determine the beach state classification (i.e., Wright and Short, 1984) given a  
454 range of hydrodynamic data and sedimentological data. The study specifically found utility in  
455 the uncertainty of prediction, an intrinsic part of the Bayesian estimation that was used to  
456 develop probabilistic beach state predictions. At larger scales, Bayesian networks have also been  
457 used to make probabilistic predictions of coastal morphology at large scales. Gutierrez et al.,  
458 (2011) used a BN to develop shoreline change rate predictions for the US east coast based on  
459 hydrodynamic, geologic and cross shore morphology of the coastline. Plant et al. (2016)  
460 modified the BN of Gutierrez et al. (2011) to include and probabilistically predict shoreline  
461 change as well as dune height for work in the Gulf of Mexico. Interestingly, the inclusion of  
462 dune height as an input variable increases precision but predictions are not more accurate. Yates  
463 and Le Cozannet (2012) also used the approach of Gutierrez et al. (2011) to probabilistically  
464 assess future European coastal evolution (either erosion, stable or accretion) using  
465 geomorphology, geology, mean tidal range, rate of sea level rise (RSLR), and mean significant  
466 wave height. Stable coastlines are predicted with greater accuracy compared to erosive or  
467 accreting coastlines, and the authors suggest that incorporating more local behaviors may resolve  
468 this issue. Bulteau et al. (2015) use a Bayesian network to predict shoreline change on La



469 Réunion island from geomorphic setting, presence of an estuary, anthropogenic structures,  
470 RSLR, and a function of wave energy. In addition to discussing why certain inputs are more  
471 predictive than others (specifically local geomorphology), this study also specifically examines  
472 areas of misprediction, and offers insight as to the unique situations when misprediction arises.  
473 Lentz et al. (2016) use a BN to relate land cover classifications, current elevations, and expected  
474 changes in RSLR to likelihood that coastal geomorphic settings would evolve to keep up with  
475 sea-level rise or inundate. A BN approach was used specifically in this study because of its  
476 probabilistic nature. Linking large scales and small scales, Gutierrez et al. (2015) predicted  
477 decadal changes in barrier island geomorphology of Assateague Island, USA using both large-  
478 scale (e.g., shoreline change rate, distance to an inlet) and small-scale morphologic features (e.g.,  
479 dune height, beach width) in a Bayesian network.

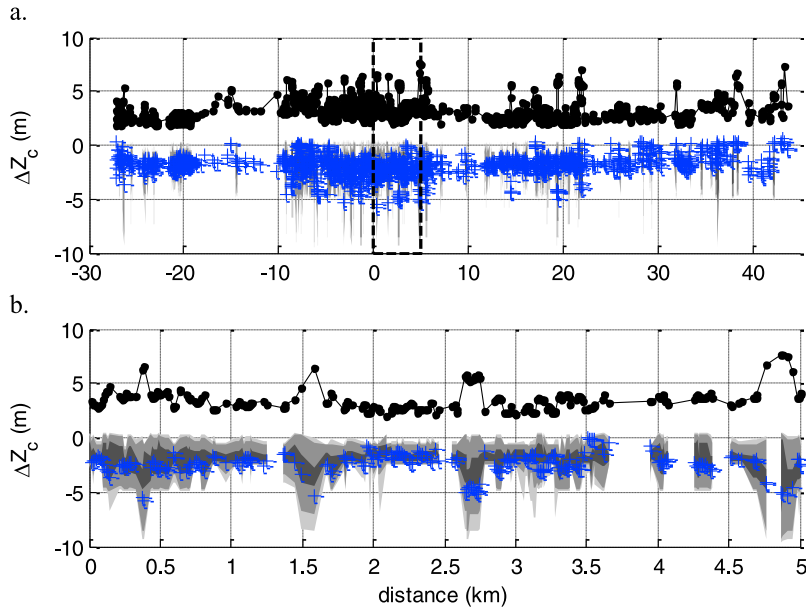
480 In addition to looking at long term shoreline changes, event scale work has also used BNs  
481 for prediction. Wilson et al. (2014) built on previous work by Lentz and Hapke (2011) to predict  
482 the beach volume changes resulting from storm events on Fire Island NY, USA with a BN.  
483 Predictions were improved in this network by including anthropogenic impacts on the beach  
484 (nourishment in this location) and adjusting the hydrodynamic inputs to the model (runup  
485 elevation vs., impact hours). This highlights the potential role of using several different inputs  
486 that may be viewed as quantifying the same process (wave-driven erosion), but may vary in  
487 correlation with the desired output (beach volume change). Beuzen et al. (2017) compared the  
488 use a BNs of different size (2 vs. 3 input nodes) to predict shoreline retreat as a consequence of  
489 storm events and preexisting beach characteristics (state, slope, width) at Collaroy-Narrabeen  
490 Beach in SE Australia. Building on this work, Beuzen et al. (2018) examined the use of BNs as  
491 both predictive tools (high performance on testing data) and descriptive tools (high performance  
492 on training data) for storm-driven shoreline change. Beuzen et al. (2018) notes that BNs built for  
493 descriptive purposes can be used to gain insight on underlying processes that produce the data,  
494 including causality.

495 Bayesian networks have also focused on emulating process-based models of storm erosion  
496 that are particularly computationally intensive—Poelhekke et al. (2016) used a BN as an  
497 emulator for the detailed process based model XBeach (Roelvink et al. 2009). By developing a  
498 set of forcing conditions for XBeach and running the model for each forcing condition,  
499 Poelhekke et al. (2016) trained the BN to predict morphodynamic impacts (overwash depth, flow

500 velocity, and erosion) on Praia de Faro, Portugal, a developed barrier island. The goal of the  
501 work is to develop a quick method to emulate the XBeach for use in an early warning system.  
502 Plomaritis et al. (2017) extended the work of Poelhekke et al. (2016) and a BN on disaster risk  
503 reduction (Jager et al., 2017) to assess the impact of risk reduction measures on morphodynamic  
504 impacts on the Ria Formosa Barrier system of Portugal. Again the BN served as mechanism to  
505 emulate process based model runs.

### 506 **4.3 Dune Erosion**

507 BN studies of coastal dune erosion from storm events have also found value in explicitly  
508 developing probabilistic prediction —Plant and Stockdon (2012) used a BN to predict dune crest  
509 elevation changes, dune crest position change, and shoreline position change as a function of  
510 dune base elevation, storm induced mean water level, and storm induced run-up (Figure 4).  
511 Observations of dune erosion do not always match predictions perfectly, but do fall within the  
512 confidence intervals of the probabilistic method — this highlights the utility of probabilistic  
513 predictions toward enhancing prediction accuracy and certainty. Palmsten et al. (2014) used the  
514 network from Plant and Stockdon (2012) as well as a simplified model structure to develop  
515 probabilistic predictions of dune position change along the Gold Coast in Queensland, Australia.  
516 Of note in this study is the attempt to use the trained model from Plant and Stockdon (2012),  
517 with no modifications or additional training, for a new site — prediction was not skillful with  
518 this model, however the ability for ML models to be generalized and extrapolated to new sites is  
519 an important test for any coastal ML models. den Heijer et al (2012) performed a similar test  
520 using a Bayesian network that was designed to emulate an existing volumetric dune erosion  
521 model. The trained model was not able to successfully extrapolate beyond the range of the  
522 training data.



523

524 *Figure 4: Figure from Plant and Stockdon (2012), who used a Bayesian network to make*  
525 *predictions of foredune crest elevation change ( $\Delta Z_c$ ). Initial dune height is shown in black,*  
526 *observations are blue +, and predictions from the BN are shown as shaded area ranging from*  
527 *50% (dark) to 95% (light) confidence interval.*

528

#### 529 **4.4 Cliffs and Rocky Coastlines**

530 Much of the previous work has focused on low-sloped sandy coastlines, though there has  
531 been work on rocky coastlines. Dickson and Perry (2015) used several regression tree  
532 approaches to identify the controls on coastal cliff landsliding (e.g., distance to fault, bedding  
533 dip, aspect, etc.). Multiple methods converged on the same two controlling variables, a benefit  
534 when comparing multiple ML methods. Hapke and Plant (2010) used a BN to develop a  
535 relationship between short term cliff erosion rate of rocky coastlines of the southern California,  
536 US, and underlying geology, cliff height, cliff slope, and a metric based on hours the cliff is  
537 subject to wave attack, and long term erosion rate of the cliffs.

#### 538 **4.5 Wave Ripples**

539 Shifting from the coastline to smaller scale morphology, Yan et al. (2008) built an ANN to  
540 predict wave generated ripple size (length and height) based on sedimentological and  
541 hydrodynamic conditions. Data were from both field and lab studies, and the ANN results were  
542 compared to four other empirical models. The ANN results provide more accurate predictions

543 based on 3 statistical measures (scatter index, correlation coefficient, and mean geometric  
544 deviation) than the empirical models. Also studying wave-generated ripple geometry, Goldstein  
545 et al., (2013) used a GP routine to construct an equation for wave generated ripple height,  
546 wavelength, and steepness using sediment grain size and near bed orbital excursion. The new  
547 machine learning scheme produced more accurate predictions compared to traditional predictors.  
548 This predictor was ultimately used as a component within a larger numerical model (Goldstein et  
549 al., 2014).

#### 550 **4.6 Flora and Fauna**

551 We can find very few coastal morphodynamic/morphology studies that focus on machine  
552 learning with flora and fauna. Coco et al. (2011) used ANN to predict the change in elevation of  
553 intertidal flats based based on seagrass (*Zostera noltii*) shoot density, leaf length, leaf area index,  
554 wave height, wind speed, sand content of the bed sediment, and sediment dry density. Gieder et  
555 al. (2014) used a Bayesian network to predict piping plover (*Charadrius melodus*) nest presence  
556 based on a Bayesian network that included barrier island morphology. Building on this work,  
557 Zeigler et al. (2017) used data collected from a variety of practitioners using a phone application  
558 to provide data to a BN to predict piping plover (*Charadrius melodus*) nest location. These three  
559 studies highlight a potential future direction for coupling coastal morphodynamics to the  
560 dynamics of flora and fauna.

#### 561 **4.7 Detection of Bars and Shoreline in images**

562 Detection of morphological features from video images has also employed regression-  
563 based ML. Kingston et al. (2000) used an ANN to model the difference between sandbar position  
564 and video intensity maxima with additional inputs of wave height and tide level. Additionally,  
565 the model developed by Kingston et al. (2000) showed success against other methods (Plant et  
566 al., 2007). Related work has focused on detecting the shoreline in video observations with a  
567 variety of ANN architectures (Alvarez-Ellacuria et al., 2011; Rigos et al., 2016).

568

### 569 **5. Hybrid ML Morphodynamic models**

570 Machine learning methods do not need to operate alone, and can be linked with  
571 morphodynamic models to create what we refer to as ‘hybrid’ models, after the atmosphere and  
572 ocean models by Krasnapolsky and Fox-Rabinovitz (2007), and Krasnapolsky (2013). There are  
573 several reasons for a hybrid models (Goldstein and Coco, 2015): ML components can serve as

574 ‘emulations’ of complex routines or equations to speed up the computational process; data-  
575 driven parameterizations can serve as model components when parameterizations have ample  
576 data but no single optimal expression — perhaps there are multiple competing formulations;  
577 more data might be anticipated in the near term future, and the parameterizations might be  
578 ‘volatile’, subject to change as new data is collected; lastly, hybrid models offer a degree of  
579 specificity to a model. Adding a ML predictor is way of incorporating a bespoke prediction  
580 scheme, which can be useful for modeling a specific setting where data was collected.

581 Three coastal morphodynamic models have combined genetic programming routines to aid  
582 in various aspects of modeling. Goldstein et al., (2014) incorporated a GP derived suspended  
583 sediment reference concentration predictor and equilibrium wave orbital ripple morphology  
584 predictor (Goldstein et al., 2013) into a model of inner shelf sorted bedforms. The model  
585 previously had been built using theoretical and empirical parameterizations of these processes,  
586 but data from inner shelf sorted bedforms was used to develop new parameterizations and  
587 produce a refined model. The goal of the modeling work was to add more specificity to the  
588 process parameterizations in settings with mixed grain sizes. Limber et al. (2014) and Limber  
589 and Murray (2014) used a GP derived expression as a component in nonlinear dynamical system  
590 model for rocky coastline evolution. The GP routine was used to develop an expression that  
591 emulated the output of a wave ray tracing model, thus summarizing the wave model results into a  
592 single smooth continuous equation amenable to further numerical work and phase plane analysis.  
593 Finally, Goldstein and Moore (2016) developed a nonlinear dynamical model of coastal dunes  
594 subject to storms by combined an empirical formulation of coastal foredune growth with a  
595 parameterization for dune erosion built using a GP from data reported in the literature. The GP  
596 routine was used to fit a smooth continuous equation to a set of data to facilitate numerical  
597 analysis.

598 Bayesian networks have been used as subcomponents for a variety of coastal models. Plant  
599 et al. (2014) used a Bayesian network to estimate overwash probability of a berm from  
600 hydrodynamic and wind conditions. This overwash probability was linked to a (non-BN) model  
601 of berm morphology. At larger space and time scales, both Passeri et al. (2016) and Bilskie et al.  
602 (2016) use the BN of Plant et al. (2016) — itself an extension of Gutierrez et al. (2014) —as a  
603 model component to predict century-scale shoreline change and dune height change as a function  
604 of SLR scenarios and geological constraints for the Gulf of Mexico. Bilskie et al. (2016) used the

605 BN as a component in a model of Hurricane impacts under different SLR scenarios while Passeri  
606 et al. (2016) used the BN as a component of a model to simulate tidal hydrodynamics under SLR  
607 scenarios. Both Passeri et al. (2016) and Bilskie et al. (2016) mention that the Bayesian network  
608 was used because it is computationally efficient for the long time and large space scales that  
609 were modeled. Passeri et al. (2016) also discuss limitations to the BN component — the lack of  
610 historic data to train the BN limited its use in bays and estuaries, and large scale barrier island  
611 processes such as rollover of back barrier shoreline migration and nourishment also were not  
612 encoded in the BN (but included as rules in the larger model). van Verseveld et al. (2015) used  
613 the process based model XBeach to simulate a storm event impacts on a developed barrier island.  
614 Hydrodynamic and morphodynamic output from the model was used as input for a Bayesian  
615 network, which predicted the damage to buildings. van Verseveld et al. (2015) notes that BN  
616 require a significant amount of data, but much was available for this research question, enabling  
617 a data driven approach. The use of multiple inputs (flooding, scour, wave height) and the  
618 probabilistic nature of the BN were also advantages in this research.

619 Many coastal morphodynamic models have several free parameters that must be ‘tuned’ for  
620 a given field site or use case (e.g., Apotsos et al., 2008, Lin and Sheng, 2017; Murray et al.,  
621 2016; Pinsky et al., 2013; Plant and Holland, 2011; Stephens et al., 2011, Stockdon et al., 2014).  
622 When the number of free parameters is large and potentially interrelated, machine learning can  
623 be used to find optimal parameters. Knaapen and Hulscher (2002) developed a model for sand  
624 wave growth and saturated morphology, with best-fit model parameters are found using of a GA  
625 routine. Ruessink (2005) tuned nearshore model parameters using a genetic algorithm coupled to  
626 a local optimization routine (downhill simplex). Komurcu et al. (2008) used a GA to determine  
627 the values for coefficients in two highly nonlinear functions that predict experimentally produced  
628 bar geometry based on the wave height, wave period, bed slope and grain size.

629 We have so far discussed ML components that are internal to morphodynamic models  
630 (hybrid models) and ML that is used to tune models. ML can also be used to analyze model  
631 results and provide insight into model output. Lazarus et al. (2011) use nonlinear forecasting to  
632 quantify the nonlinearity of timeseries output generated by a model of human – coastline  
633 interaction.

## 634 **6. Discussion**

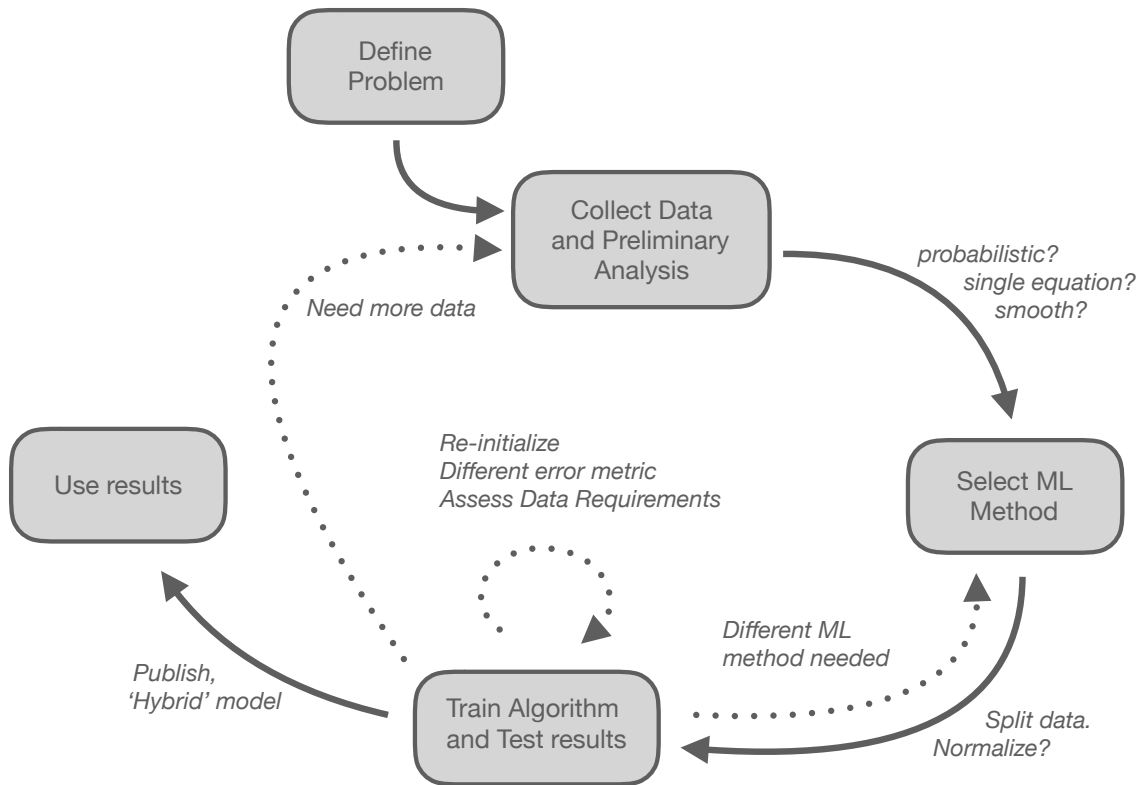
635 We have identified a wide range of ML applications to sediment transport, morphologic,

636 and hybrid coastal prediction problems. This provokes questions, such as how to select an  
637 appropriate ML method to suit a particular problem, what sort of comparisons of ML can be  
638 made to more standard approaches, what are the common principles that can be applied to these  
639 applications, and where does this lead. These topics are discussed below.

#### 640 **6.1 Which ML technique and how much data?**

641 We have reviewed commonly used ML techniques in the field of coastal morphodynamics,  
642 and investigated their applications across a range of scales. Figure 5 is a schematic representation  
643 of our workflow. Some of the steps in this workflow are decisions (selecting an algorithm,  
644 determining if more data is needed, etc.) and may require significant investment of thought. With  
645 many possible tools, we are often asked about the best ML routine to use for a given type of  
646 problem. Unfortunately in most instances we cannot definitively answer this question a priori,  
647 but we can offer some insights to readers wondering where to start. A researcher may have a  
648 preference for a given learning routine based on the type or quantity of data, an intuition  
649 regarding what worked best on a similar problem, or the form of a desired outcome. Guiding  
650 questions can be used to determine the ML routine for a given problem: If probabilistic answers  
651 are required, researchers could focus on Bayesian networks. If multiple free parameters exist for  
652 a fixed, immutable equation, then a GA can be employed. If a specific smooth equation is  
653 needed (e.g., to be used in an analytical model), an ANN or GP can be used. If functional form  
654 and input dependencies are needed, a solution from GP can be attempted. Ambiguity over  
655 picking a ML approach for a given dataset highlights the need for a larger empirical study where  
656 many ML approaches are each attempted on an array of problems to determine (empirically) if  
657 there are optimal techniques for a given research question. Examples of this approach can be  
658 seen in other disciplines — work by Olson et al. (2017) and Hansen et al. (2013). Below (section  
659 6.3) we offer a set of best practices for coastal researchers who aspire to make their ML results  
660 usable for ML comparison.

661

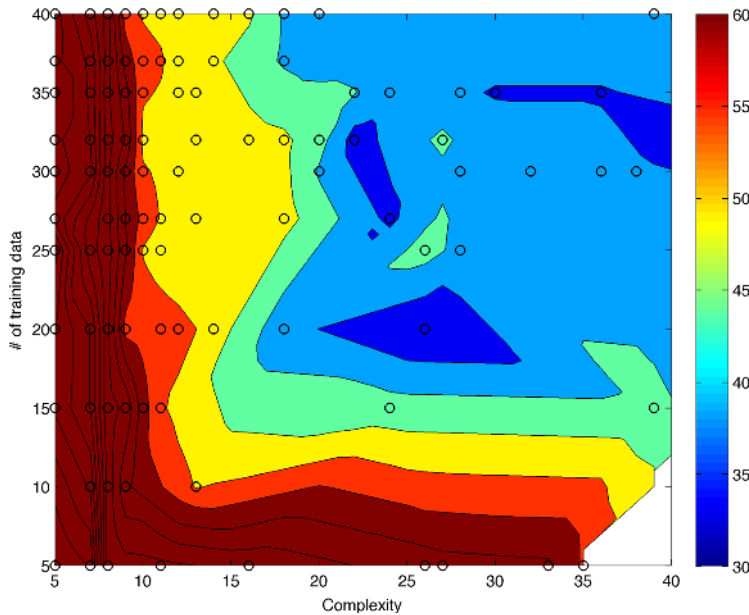


662  
663 *Figure 5: A schematic of our workflow. Work starts with ‘Define Problem’—new or published*  
664 *data — and progresses by following the filled arrows. Dotted arrows represent escapes from this*  
665 *workflow for various reasons*

666  
667 It remains unclear how much data a researcher needs to perform meaningful ML analysis  
668 and produce a useful predictor. For example, Beuzen et al. (2017) investigated the amount of  
669 data needed to successfully parameterize a BN of shoreline change as a function of storm events,  
670 and suggested that the number likely depends on network complexity (degrees of freedom, or  
671 independent variables) and the signal ‘clarity’, likely a measure of signal to noise in the dataset.  
672 Goldstein and Coco (2014) and Tinoco et al. (2015) investigated the impact of adding data for  
673 GP prediction, finding that very little data was needed to train the model, and more data might  
674 actually provide degraded prediction for low-complexity models (Figure 6). It is unclear if this  
675 relationship holds for other machine learning routines or prediction tasks, but could point the  
676 way in minimizing training data and maximizing the data used to test the resultant ML derived  
677 predictor. Similar to ‘which algorithm’, we expect that empirical guidance is most valuable, and



678 further studies might shed light on how much data is really needed to make optimal predictions  
679 for a given problem, with a given prediction technique. Providing justification for a given  
680 amount of data, or a given ML technique — especially quantitative justification — is particularly  
681 valuable for future researchers to determine what techniques and data quantities work for a given  
682 problem, and what can be learned (or what we are unable to learn) using a given approach and  
683 given set of data.



684  
685 *Figure 6: A plot from Goldstein and Coco (2014) that displays GP solutions (open circles) of*  
686 *differing complexity developed by changing the size of the training data. The colormap is a*  
687 *measure of error, and suggests that continuing to increase the size of training data does not*  
688 *always yield decreases in error of the final solution.*

689  
690 Beyond the amount of data needed for a ML analysis, all machine learning methods require  
691 that data be split into discrete parts — with one part of the data used for developing the model  
692 (‘training data’), while the remaining data are reserved outside the algorithm to test the  
693 developed model (‘testing data’). The most important aspect of the data splitting is that users do  
694 not test the learned predictor with the training data (e.g., Domingos, 2012) — the testing data  
695 should not be seen by the ML algorithm as it is trained, testing must occur using a new, unseen  
696 portion of the data. There are enhancements beyond this simple two-part split (i.e., it is common  
697 to use additional subsets in the training process as ‘validation’ to limit overfitting), but for this

698 moment it is important that some portion of the data is used in the model building phase, and  
699 some portion of data is used in the model testing phase.

700 In many of the studies reviewed in this paper the data are split randomly. While random  
701 selection can be used, there are many methods to select a data for training and testing that are  
702 advantageous for developing generalizable predictor. For example, it is often easier to acquire  
703 data for coastal studies in controlled lab settings under weak forcing, and there may be few data  
704 points in field conditions (with complexities such as mixed grain sizes, irregular waves, etc.) or  
705 under extreme forcing (i.e., storms). Extracting training data by randomly sampling a full dataset  
706 may omit the critical information from the training data, for example extreme conditions (e.g.,  
707 Passarella et al., 2018). There are many intentional sampling strategies from splitting data into  
708 training and testing that have not seen wide adoption in coastal community — we believe that  
709 adoption of these techniques will lead to more generalizable ML predictors. For examples of data  
710 selection routines, see Galelli et al. (2014) on evaluating input value selection routines, and  
711 previous work by Bowden et al. (2002), Bowden et al. (2012), Camus et al. (2011a; 2011b), May  
712 et al. (2010), Tinoco et al. (2015), Splinter et al. (2013).

713

## 714 **6.2 When does ML perform better than more traditional methods?**

715 As we state in the sediment transport modeling section, the ML studies we review often  
716 outperform more traditional prediction schemes that are based on derivations from ‘first  
717 principles’ (i.e., conservation laws) and schemes based on more classic curve fitting techniques  
718 (e.g., linear regression). It remains unclear why this is the case. One possibility is that we are  
719 only aware of published ML studies, which bias us into believing ML provides only positive  
720 results. Or perhaps developing a new ML predictor might outperform a non-ML predictor  
721 because of flexibility — many ML routines can develop predictors that do not conform to a set  
722 basis function. For example, more traditional regression techniques work well when data  
723 conforms to a set functional form (a line, or a curve) and obeys the many generalizing  
724 assumptions (i.e., normally distributed). A researcher therefore performs dimensional reduction  
725 or transformation to first get data into this functional form, then fits the data with a known basis  
726 function (a line, a parabola, etc.). ML techniques offer more flexibility because the basis function  
727 for many techniques is highly adaptable — i.e., a neural network, which has many free  
728 parameters and can be trained to fit lines or curves.

729 We have discussed many of the studies that develop ML predictors for the same  
730 phenomena (e.g., suspended sediment transport flux, bar geometry, ripple wavelength). Even  
731 when multiple studies examine the same research question, the comparison between ML  
732 predictors is often impossible because each study uses different datasets, the method to split and  
733 testing data is often not clear, or information regarding the final predictor is not provided (e.g.,  
734 for cases where ANN are used, the weights and biases may not be provided). Therefore it is  
735 difficult to truly compare ML methods and the resulting predictors. For instances, it is difficult to  
736 compare the ripple predictors developed by Yan et al. (2008) with a ANN and the ripple  
737 predictors developed by Goldstein et al. (2013) with a GP. After reviewing the work of ML in  
738 coastal morphodynamics and sediment transport we are left instead with the knowledge that  
739 the studies reviewed here can be understood as a ‘proof of concept’ for ML being able to develop  
740 accurate predictions for a given dataset. In this way, we suggest that ML might point the way  
741 toward the being valuable for developing bespoke prediction routines for a given site, a given  
742 dataset, or a given purpose. In some cases, more accurate prediction might be needed because of  
743 a specific research question or prediction task. An example is work by Goldstein et al. (2014),  
744 where a near bed suspended sediment reference concentration prediction scheme was developed  
745 from data collected in fields of sorted bedforms, and then used in a model of sorted bedforms to  
746 predict sorted bedform dynamics.

### 747 **6.3 A set of practices for Coastal ML research**

748 If researchers using ML aspire to make their results reproducible, transferable, and useful  
749 to future researchers (e.g., intercomparison projects, helping to determine which algorithm is  
750 best, or how much data is needed), we offer a set of practices here that would aid in this goal. 1)  
751 Provide the data with the paper, or link to an open archive. 2) Unequivocally state the degree to  
752 which the training data and testing data are separate, that the testing data was not seen by the  
753 machine learning algorithm, and that the training data was not used to test the success of the  
754 developed model. 3) Clearly describe the technique used to split the data into training and  
755 testing, the percentage of data used for each group, and if possible, the actual data split into  
756 groups. 4) Report the final model in its entirety (e.g., weights, biases and architecture for ANN;  
757 binary splits for a tree model, etc.). 5) Define the metrics that are used to test the models and  
758 define levels required to be successful. 6) Compare results to other models to provide  
759 benchmarks for improvements and the relative value of ML vs. theoretically developed models.

760 7) Compare ML model to newly collected data sets as a test to determine whether there are  
761 sufficient data for a particular model and whether the model is locally or generally applicable.

762 The goal with this set of best practices is to steer ML papers toward being usable and  
763 reproducible by the community. We understand that for a variety of reasons, these practices may  
764 not be possible under all circumstances, but if authors would like their ML work be built upon,  
765 tested, and refined, these practices aid in that goal. Very few studies that we review adhere to all  
766 of these practices. However we offer this guidance here to help advance the understanding and  
767 reuse of ML research in the coastal community, and we hope that all future work is performed  
768 with an eye toward reuse and reproducibility by others.

769

#### 770 **6.4 Future directions**

771 Data-driven research relies on the existence of data. Beyond the collection and ad hoc  
772 sharing of data, the trend in publication of data is a major factor in the continued adoption of  
773 data-driven science. Data publication is enabled by the existence of repositories — Figshare,  
774 Zenodo, Data Dryad, Pangaea, and others identified by the re3data.org project (Pampel et al.,  
775 2013) — as well as data journals, publications that focus exclusively on descriptions of the data  
776 (from collection to access) such as Earth System Science Data (Pfeiffenberger and Carlson,  
777 2011), Earth and Space Science (Hanson, 2014), Geoscience Data (Allan, 2014), and Scientific  
778 Data (Scientific Data, 2014). The continued collection and release of data will enable more data-  
779 driven work to occur in coastal settings.

780 In addition to data, ML research relies on ML algorithms and techniques. We have only  
781 discussed ML techniques that have been used in coastal settings, but this in no way is an  
782 exhaustive list of techniques. First, many common techniques — such as Support Vector  
783 Machines — have seen only minimal usage in coastal morphodynamics problems; though they  
784 have been used for coastal classification routines (Hoonhout et al., 2015) and in oceanographic  
785 contexts (Li et al., 2013). Second, many newer techniques may not yet have been applied to  
786 coastal research, such as recent advances in ANN architecture and training (e.g., Deep Learning;  
787 LeCun et al., 2015), or newer probabilistic techniques (Ghahramani, 2015). There is a world of  
788 new algorithms and techniques that can be brought over from the ML community — researchers  
789 might find it profitable to look for new techniques within the ML literature to make sure we are  
790 not missing out on the revolutionary advances in data-driven tools.

791 Finally, ML learning techniques can be thought of as new additions to the coastal  
792 researchers bag of tools — providing new insights during data analysis (e.g. Tinoco et al., 2015;  
793 Beuzen et al., 2018). ML techniques could be taught alongside other more common data analysis  
794 techniques (e.g., Fourier transforms, wavelets; Zdeborova, 2017) since ultimately the goal is the  
795 same with any of these tools and techniques — to find and extract new knowledge or insight  
796 from data.

797 In light of our general requests to continue publishing data for reuse, the continued  
798 adoption of new algorithms, and to teach these modern methods to students, we also see three  
799 specific areas for growth in Coastal ML research. First, to provide guidance for which methods  
800 perform best for which problems, a more structured comparison projects between ML techniques  
801 using the identical data set (and data split) is required. Comparative work of this nature will help  
802 all researchers decide which ML has a high chance of success for a given problem. This may also  
803 help us to further understand a given coastal problem. Second, we identify an opportunity to use  
804 ML on timeseries, especially when systems may have memory and/or storage effects. The work  
805 of Pape et al (2010) is a rare example, however new advances in neural network architecture  
806 (i.e., long short-term memory) has the potential to allow for more accurate time series prediction  
807 even when systems have strong autocorrelation, thresholds, and memory dynamic (e.g.,  
808 shorelines, bars, bedforms, etc.). See Kratzert et al. (2018) as an example of the power of long  
809 short-term memory networks in hydrological time series prediction. Third, uncertainty derived  
810 from ML can further be incorporated into models. A clear possibility is to use probabilistic ML  
811 based predictions (from Bayesian networks, or other novel techniques such as Gaussian  
812 Processes) creatively in numerical models. An example is to use the probabilistic nature of these  
813 predictors as ‘stochastic parameterizations’ (e.g., Berner et al., 2017), whereby some aspect of an  
814 otherwise deterministic numerical model is made probabilistic, and models may then be able to  
815 generate ensemble predictions using identical forcing and initial conditions.

816

817 **Acknowledgements:**

818 EBG gratefully acknowledges the support of UoA through a PBRF grant. GC funded by a GNS-  
819 Hazard Platform grant (contract 3710440). We are grateful for a review by Chris Sherwood at  
820 USGS.

821

822 **References:**

- 823 Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R., Salmerón, A., 2011. Bayesian networks  
824 in environmental modelling. *Environmental Modelling & Software*, 26(12), 1376-1388.
- 825 Allan, R., 2014. Geoscience data. *Geosci. Data J.*, 1: 1. doi:10.1002/gdj3.3
- 826 Álvarez-Ellacuría, A., Orfila, A., Gómez-Pujol, L., Simarro, G., Obregón, N., 2011. Decoupling  
827 spatial and temporal patterns in short-term beach shoreline response to wave climate.  
828 *Geomorphology*, 128: 199-208.
- 829 Anderson, C., 2008. The end of theory: The data deluge makes the scientific method  
830 obsolete. *Wired Magazine*, 16(7), 16-07.
- 831 Apotsos, A., Raubenheimer, B., Elgar, S., Guza, R. T., 2008. Testing and calibrating parametric  
832 wave transformation models on natural beaches, *Coastal Engineering*, 55, 224-235.
- 833 Aubrey, D. G., 1979. Seasonal patterns of onshore/offshore sediment movement, *J. Geophys.*  
834 *Res.*, 84, 6347–6354.
- 835 Babovic, V., Keijzer, M., 2000. Genetic programming as a model induction engine. *Journal of*  
836 *Hydroinformatics*, 2(1), 35-60.
- 837 Berner, J., Achatz, U., Batte, L., Bengtsson, L., Cámara, A. D. L., Christensen, H. M., ... &  
838 Franzke, C. L. (2017). Stochastic parameterization: Toward a new view of weather and  
839 climate models. *Bulletin of the American Meteorological Society*, 98(3), 565-588.
- 840 Beuzen, T., Splinter, K. D., Turner, I. L., Harley, M. D., Marshall, L., 2017. Predicting storm  
841 erosion on sandy coastlines using a Bayesian network. *Australasian Coasts & Ports*  
842 *2017: Working with Nature*, 102.
- 843 Beuzen, T., Splinter, K. D., Marshall, L. A., Turner, I. L., Harley, M. D., & Palmsten, M. L.  
844 (2018). Bayesian Networks in coastal engineering: Distinguishing descriptive and  
845 predictive applications. *Coastal Engineering*, 135, 16-30.
- 846 Bilskie, M. V., Hagen, S. C., Alizad, K., Medeiros, S. C., Passeri, D. L., Needham, H. F., Cox,  
847 A., 2016. Dynamic simulation and numerical analysis of hurricane storm surge under sea  
848 level rise with geomorphologic changes along the northern Gulf of Mexico, *Earth's*  
849 *Future*, 4, 177–193, doi:10.1002/2015EF000347.
- 850 Bolaños, R., Souza, A., 2010. Measuring hydrodynamics and sediment transport processes in the  
851 *Dee Estuary*, *Earth Syst. Sci. Data*, 2, 157-165, doi:10.5194/essd-2-157-2010.

- 852 Bowden, G. J., Maier, H. R., Dandy, G. C., 2002. Optimal division of data for neural network  
853 models in water resources applications. *Water Resources Research*, 38(2).
- 854 Bowden, G. J., Maier, H. R., Dandy, G. C., 2012. Real-time deployment of artificial neural  
855 network forecasting models: Understanding the range of applicability. *Water Resources*  
856 *Research*, 48(10).
- 857 Bulteau, T., Baills, A., Petitjean, L., Garcin, M., Palanisamy, H., Le Cozannet, G., 2015. Gaining  
858 insight into regional coastal changes on La Réunion island through a Bayesian data  
859 mining approach. *Geomorphology*, 228, 134-146.
- 860 Camus, P., Mendez, F. J., Medina, R., Cofiño, A. S., 2011a. Analysis of clustering and selection  
861 algorithms for the study of multivariate wave climate. *Coastal Engineering*, 58(6), 453-  
862 462.
- 863 Camus, P., Cofiño, A. S., Mendez, F. J., Medina, R., 2011b. Multivariate wave climate using  
864 self-organizing maps. *Journal of Atmospheric and Oceanic technology*, 28(11), 1554-  
865 1568.
- 866 Casdagli, M., 1989. Nonlinear prediction of chaotic time series. *Physica D: Nonlinear*  
867 *Phenomena*, 35(3), 335-356.
- 868 Casdagli, M., 1992. A dynamical systems approach to modeling input-output systems. In *Santa*  
869 *Fe Institute Studies in the Sciences of Complexity- Proceedings* Vol. 12, pp. 265-265.  
870 Addison-Wesley Publishing C.
- 871 Chang, C. W., Ushio, M., Hsieh, C. H., 2017. Empirical dynamic modeling for  
872 beginners. *Ecological Research*, 32(6), 785-796.
- 873 Charniak, E., 1991. Bayesian networks without tears, *AI Magazine*, 12(4), 50-63.
- 874 Chau, K., 2006. A review on the integration of artificial intelligence into coastal  
875 modeling. *Journal of Environmental Management*, 80(1), 47-57.
- 876 Chen, S. H., Pollino, C. A., 2012. Good practice in Bayesian network modelling. *Environmental*  
877 *Modelling & Software*, 37, 134-145.
- 878 Coco, G., Ganthly, F., Sottolichio, A., Verney, R., 2011. The use of Artificial Neural Networks  
879 to predict intertidal sedimentation and unravel vegetation effects Proceedings of the 7th  
880 IAHR Symposium on River, Coastal, and Estuarine Morphodynamics, Beijing, China,
- 881 Cohen, I. B., Whitman, A., Budenz, J., 2016. The Principia: The Authoritative Translation and  
882 Guide: Mathematical Principles of Natural Philosophy.

- 883 Cooper, G. E., Herskovits, E., 1992. A Bayesian Method for the Induction of Probabilistic  
884 Networks from Data, *Machine Learning*, 9(309-347).
- 885 Coveney, P. V., Dougherty, E. R., Highfield, R. R., 2016. Big data need big theory too. *Phil.*  
886 *Trans. R. Soc. A*, 374(2080), 20160153.
- 887 Crutchfield, J. P., 2014. The dreams of theory. *Wiley Interdisciplinary Reviews: Computational*  
888 *Statistics*, 6(2), 75-79.
- 889 D'Ambrosio, D., Spataro, W., Rongo, R., Iovine, G., 2013, Genetic algorithms, optimization,  
890 and evolutionary modeling. In: Shroder, J. (Editor in Chief), Baas, A. C. W. (Ed. ),  
891 *Quantitative Modeling of Geomorphology*. Academic Press, San Diego, CA, vol. 2, pp.  
892 74–97.
- 893 De'ath, G., Fabricius, K. E., 2000. Classification and regression trees: a powerful yet simple  
894 technique for ecological data analysis. *Ecology*, 81(11), 3178-3192.
- 895 De'Ath, G., 2007. Boosted trees for ecological modeling and prediction. *Ecology*, 88(1), 243-  
896 251.
- 897 Demirci, M., Üneş, F., Aköz, M. S., 2015. Prediction of cross-shore sandbar volumes using  
898 neural network approach. *Journal of Marine Science and Technology*, 20(1), 171-179.
- 899 den Heijer, C. K., Knipping, D. T., Plant, N. G., de Vries, J. S. V. T., Baart, F., & van Gelder, P.  
900 H. (2012). Impact assessment of extreme storm events using a Bayesian network. *Coastal*  
901 *Engineering Proceedings*, 1(33), 4.
- 902 Dickson, M. E., Perry, G. L., 2016. Identifying the controls on coastal cliff landslides using  
903 machine-learning approaches. *Environmental Modelling & Software*, 76, 117-127.
- 904 Dietterich, T., 1995. Overfitting and undercomputing in machine learning. *ACM computing*  
905 *surveys (CSUR)*, 27(3), 326-327.
- 906 Domingos, P., 2012. A few useful things to know about machine learning. *Communications of*  
907 *the ACM*, 55(10), 78-87.
- 908 Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. *Journal*  
909 *of Animal Ecology* 77, 802–813.
- 910 Farmer, J. D., Sidorowich, J. J., 1987. Predicting chaotic time series. *Physical review letters*,  
911 59(8), 845
- 912 Galelli, S., Humphrey, G. B., Maier, H. R., Castelletti, A., Dandy, G. C., Gibbs, M. S., 2014. An  
913 evaluation framework for input variable selection algorithms for environmental data-



- 914 driven models. *Environmental modelling & software*, 62, 33-51.
- 915 Garel, E., Ferreira, Ó., 2015, Multi-year high-frequency physical and environmental observations  
916 at the Guadiana Estuary, *Earth Syst. Sci. Data*, 7, 299-309, doi:10.5194/essd-7-299-2015.
- 917 Gieder, K. D., Karpanty, S. M., Fraser, J. D., Catlin, D. H., Gutierrez, B. T., Plant, N. G.,  
918 Thieler, E. R., 2014. A Bayesian network approach to predicting nest presence of the  
919 federally-threatened piping plover (*Charadrius melodus*) using barrier island  
920 features. *Ecological modelling*, 276, 38-50.
- 921 Ghahramani, Z., 2015. Probabilistic machine learning and artificial intelligence. *Nature*,  
922 521(7553), 452-459.
- 923 Goldstein, E. B., Coco, G., Murray, A.B., 2013. Prediction of Wave Ripple Characteristics using  
924 Genetic Programming, *Continental Shelf Research*, 71, 1-15,  
925 doi:10.1016/j.csr.2013.09.020
- 926 Goldstein, E. B., Coco, G., Murray, A.B., Green, M. O., 2014. Data driven components in a  
927 model of inner shelf sorted bedforms: a new hybrid model, *Earth Surf. Dynam. Discuss.*,  
928 1, 531-569, doi:10.5194/esurfd-1-531-2013.
- 929 Goldstein, E. B., Coco, G., 2014. A machine learning approach for the prediction of settling  
930 velocity. *Water Resources Research*, 50(4), 3595-3601.
- 931 Goldstein, E. B., Coco, G., 2015. Machine learning components in deterministic models: hybrid  
932 synergy in the age of data. *Frontiers in Environmental Science*, 3, 33.
- 933 Goldstein, E. B., Moore, L. J., 2016. Stability and bistability in a one-dimensional model of  
934 coastal foredune height. *Journal of Geophysical Research: Earth Surface*, 121(5), 964-  
935 977.
- 936 Grimes, D. J., Cortale, N., Baker, K., Mcnamara, D. E., 2015. Nonlinear forecasting of intertidal  
937 shoreface evolution. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(10),  
938 103116.
- 939 Gutierrez, B. T., Plant, N. G., Pendleton, E. A., Thieler, E.R., 2014. Using a Bayesian network to  
940 predict shoreline change vulnerability to sea-level rise for the coasts of the United States,  
941 *U.S. Geol. Surv. Open File Rep. 2014 – 1083*, 26 pp., U.S. Geological Survey, Reston,  
942 Va.
- 943 Gutierrez, B. T., Plant, N. G., Thieler, E. R., Turecek, A., 2015. Using a Bayesian network to  
944 predict barrier island geomorphologic characteristics. *Journal of Geophysical Research:*

- 945 *Earth Surface*, 120(12), 2452-2475.
- 946 Gutierrez, B. T., Plant, N. G., Thieler, E. R., 2011. A Bayesian network to predict the coastal  
947 vulnerability to sea-level rise, *J. Geophys. Res.*, 116, F02009,  
948 doi:10.1029/2010JF001891.
- 949 Hansen, K., Montavon, G., Biegler, F., Fazli, S., Rupp, M., Scheffler, M., Müller, K. R., 2013.  
950 Assessment and validation of machine learning methods for predicting molecular  
951 atomization energies. *Journal of Chemical Theory and Computation*, 9(8), 3404-3419.
- 952 Hanson, B., 2014. AGU to Launch a New Open-Access Journal Spanning the Earth and Space  
953 Sciences. *Eos, Transactions American Geophysical Union*, 95(6), 56-56.
- 954 Hapke, C., Plant, N. G., 2010. Predicting coastal cliff erosion using a bayesian probabilistic  
955 model, *Mar. Geol.*, 278, 140–149, doi:10.1016/j.margeo.2010.10.001.
- 956 Hashemi, M.R., Ghadampour, Z., Neill, S.P., 2010. Using an artificial neural network to model  
957 seasonal changes in beach profiles, *Ocean Engineering*, 37, 1345-1356.
- 958 Hastie, T., Tibshirani, R., Friedman, J., 2009. *The elements of statistical learning: data mining,*  
959 *inference, and prediction*. Springer series in statistics
- 960 Hey, T., Tansley, S., Tolle, K. M., 2009. *The fourth paradigm: data-intensive scientific*  
961 *discovery* Redmond, WA: Microsoft research.
- 962 Holland, J. H., 1975. Adaptation in natural and artificial systems. An introductory analysis with  
963 application to biology, control, and artificial intelligence. *Ann Arbor, MI: University of*  
964 *Michigan Press*.
- 965 Hoonhout, B. M., Radermacher, M., Baart, F., Van der Maaten, L. J. P., 2015. An automated  
966 method for semantic classification of regions in coastal images. *Coastal*  
967 *Engineering*, 105, 1-12.
- 968 Hsieh, W. W., 2004. Nonlinear multivariate and time series analysis by neural network methods,  
969 *Rev. Geophys.*, 42, RG1003, doi:10.1029/2002RG000112.
- 970 Iglesias, G., Lopez, I., Carballo, R., Castro, A., 2009a. Headland-bay beach planform and tidal  
971 range: a neural network model. *Geomorphology* 112, 135–143.
- 972 Iglesias, G., López, I., Castro, A., Carballo, R., 2009b. Neural network modelling of planform  
973 geometry of headland-bay beaches. *Geomorphology*, 103(4), 577-587.
- 974 Iglesias, G., Diz-Lois, G., Pinto, F. T., 2010. Artificial Intelligence and headland-bay  
975 beaches. *Coastal Engineering*, 57(2), 176-183.

- 976 Jaffe, B.E., Rubin, D.M., 1996. Using nonlinear forecasting to learn the magnitude and phasing  
977 of time-varying sediment suspension in the surf zone. *Journal of Geophysical Research*  
978 101 (C6), 14,283–14,296.
- 979 Jäger, W. S., Christie, E. K., Hanea, A. M., den Heijer, C., Spencer, T., 2017. A Bayesian  
980 network approach for coastal risk analysis and decision making. *Coastal Engineering*.
- 981 Jones, N. S., Maccarone, T. J., 2013. Inference for the physical sciences. *Phil. Trans. R. Soc.*  
982 A 2013 371 20120493; DOI: 10.1098/rsta.2012.0493
- 983 Kabiri-Samani, A.R., Aghaee-Tarazjani, J., Borghei, S.M., Jeng, D.S., 2011. Application of  
984 neural networks and fuzzy logic models to long-shore sediment transport. *Applied Soft*  
985 *Computing*, 11, 2880–2887.
- 986 Kantz, H., Schreiber, T., 2004. *Nonlinear time series analysis (Vol. 7)*. Cambridge university  
987 press.
- 988 Kingston, K.S., Ruessink, B.G., van Enckevort, I.M.J., Davidson, M.A., 2000. Artificial neural  
989 network correction of remotely sensed sandbar location. *Marine Geology* 169, 137–160.
- 990 Kizhisseri, A.S., Simmonds, D., Rafiq, Y., and Borthwick, M., 2005. An evolutionary  
991 computation approach to sediment transport modeling. In: *Fifth International Conference*  
992 *on Coastal Dynamics*, April 4–8, 2005, Barcelona, Spain.
- 993 Knaapen, M. A. F., Hulscher, S. J. M. H, 2002, Regeneration of sand waves after dredging,  
994 *Coast. Eng.*, 46, 277–289.
- 995 Komurcu, M. I., Tutkun, N, Ozolcer, I.H., Akpinar, A., 2008. Estimation of the beach bar  
996 parameters using the genetic algorithms, *Applied Mathematics and Computation*, 195,  
997 49-60.
- 998 Komurcu, M. I., Komur, M. A., Akpinar, A., Ozolcer, I.H., Yuksek, O., 2013. Prediction of  
999 offshore bar-shape parameters resulting by cross-shore sediment transport using artificial  
1000 neural network, *Applied Ocean Research*, 40, 74-82.
- 1001 Koza, J. R., 1992. *Genetic Programming, On the Programming of Computers by Means of*  
1002 *Natural Selection*, MIT Press, Cambridge, MA, USA.
- 1003 Kratzert F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M., (2018) Rainfall-Runoff  
1004 modelling using Long-Short-Term-Memory (LSTM) networks, *Hydrol. Earth Syst. Sci.*  
1005 *Discuss.*, <https://doi.org/10.5194/hess-2018-247>.
- 1006 Krasnopolsky, V. M., Fox-Rabinovitz, M. S., 2006. A new synergetic paradigm in environmental

- 1007 numerical modeling: Hybrid models combining deterministic and machine learning  
1008 compo- nents, *Ecol. Model.*, 191, 5–18.
- 1009 Krasnopolsky, V. M., 2013. The application of Neural Networks in the Earth Sciences, Springer.
- 1010 Lazarus, E. D., McNamara, D. E., Smith, M. D., Gopalakrishnan, S., Murray, A. B., 2011.  
1011 Emergent behavior in a coupled economic and coastline model for beach  
1012 nourishment. *Nonlinear Processes in Geophysics*, 18(6), 989-999.
- 1013 LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature*, 521(7553), 436-444.
- 1014 Lentz, E. E., Hapke, C. J., 2011. Geologic framework influences on the geomorphology of an  
1015 anthropogenically modified barrier island: assessment of dune/beach changes at Fire  
1016 Island, New York. *Geomorphology*, 126(1), 82-96.
- 1017 Lentz, E. E., Thieler, E. R., Plant, N. G., Stippa, S. R., Horton, R. M., Gesch, D. B., 2016.  
1018 Evaluation of dynamic coastal response to sea-level rise modifies inundation  
1019 likelihood. *Nature Climate Change*, 6(7), 696-700.
- 1020 Li, Z., Li, L., Song, K., Cassar, N., 2013. Estimation of phytoplankton size fractions based on  
1021 spectral features of remote sensing ocean color data, *J. Geophys. Res. Oceans*, 118,  
1022 doi:10.1002/jgrc.20137.
- 1023 Lin, S., Sheng, J., 2017. Assessing the performance of wave breaking parameterizations in  
1024 shallow waters in spectral wave models, *Ocean Modelling*, 120, 41-59.
- 1025 Limber, P. W., Brad Murray, A., Adams, P. N., Goldstein, E. B., 2014. Unraveling the dynamics  
1026 that scale cross-shore headland relief on rocky coastlines: 1. Model development. *Journal*  
1027 *of Geophysical Research: Earth Surface*, 119(4), 854-873.
- 1028 Limber, P. W., Murray, A. B., 2014. Unraveling the dynamics that scale cross-shore headland  
1029 relief on rocky coastlines: 2. Model predictions and initial tests. *Journal of Geophysical*  
1030 *Research: Earth Surface*, 119(4), 874-891.
- 1031 López, I., Aragonés, L., Villacampa, Y., Serra, J. C., 2017. Neural network for determining the  
1032 characteristic points of the bars. *Ocean Engineering*, 136, 141-151.
- 1033 Lorenz, E.N., 1969a. Atmospheric predictability as revealed by naturally occurring analogues. *J*  
1034 *Atmos Sci* 26: 636–646.
- 1035 Lorenz, E.N., 1969b. Three approaches to atmospheric predictability. *Bull Am Meteorol* 50:  
1036 345–349.

- 1037 Loureiro, C., Ferrera, O., Cooper, J.A., 2013. Applicability of parametric beach morphodynamic  
1038 state classification on embayed beaches, *Mar. Geol.*, 346, 153–164.
- 1039 Luijendijk, A., Hagenaars, G., Ranasinghe, R., Baart, F., Donchyts, G., & Aarninkhof, S. (2018).  
1040 The State of the World's Beaches. *Scientific reports*, 8. Article number: 6641,  
1041 [10.1038/s41598-018-24630-6](https://doi.org/10.1038/s41598-018-24630-6)
- 1042 Mafi, S., Yeganeh-Bakhtiary, A., Kazeminezhad, M.H., 2013. Prediction formula for longshore  
1043 sediment transport rate with M5' algorithm In: Mafi, S., Yeganeh-Bakhtiary, A. and  
1044 Kazeminezhad, M.H. Proceedings 12<sup>th</sup> International Coastal Symposium (Plymouth,  
1045 England), *Journal of Coastal Research*, Special Issue No. 65, pp. 2149-2154, ISSN 0749-  
1046 0208.
- 1047 Malakoff, D., 1999. Bayes Offers a 'New' Way to Make Sense of Numbers, *Science*, 286(5444),  
1048 1460-1464.
- 1049 May, R. J., Maier, H. R., Dandy, G. C., 2010. Data splitting for artificial neural networks using  
1050 SOM-based stratified sampling. *Neural Networks*, 23(2), 283-294.
- 1051 Mitchell, M., 1995. Genetic algorithms: An overview. *Complexity*, 1(1), 31-39.
- 1052 Mitchell, M., 1998. *An introduction to genetic algorithms*. MIT press.
- 1053 Múnera, S., Osorio, A. F., Velásquez, J. D., 2014. Data-based methods and algorithms for the  
1054 analysis of sandbar behavior with exogenous variables. *Computers & Geosciences*, 72,  
1055 134-146.
- 1056 Murray, A.B., Goldstein, E.B., Coco, G., 2014a. Cause and Effect in Geomorphic Systems:  
1057 Complex-Systems Perspectives, *Geomorphology*, 219, 1-9.
- 1058 Murray, A.B., Coco, G., Goldstein, E. B., 2014b. The Shape of Patterns to Come: From Initial  
1059 Formation to Long-Term Evolution, *Earth Surface Processes and Landforms*,  
1060 doi: 10.1002/esp.3487DOI: 10.1002/esp.3487
- 1061 Murray, A.B., Gasparini, N.M, Goldstein, E.B., van der Wegen, M., 2016, Uncertainty  
1062 quantification in modeling earth surface processes: more applicable for some types of  
1063 models than for others. *Computers & Geosciences*,  
1064 <http://doi.org/10.1016/j.cageo.2016.02.008>
- 1065 Nelson, T. R., Voulgaris, G., Traykovski, P., 2013. Predicting wave-induced ripple equilibrium  
1066 geometry. *Journal of Geophysical Research: Oceans*, 118(6), 3202-3220.

- 1067 Nylén, T., Hellemaa, P., Luoto, M., 2015. Determinants of sediment properties and organic  
1068 matter in beach and dune environments based on boosted regression trees. *Earth Surface*  
1069 *Processes and Landforms*, 40(9), 1137-1145.
- 1070 O'Neill, M., Vanneschi, L., Gustafson, S., Banzhaf, W., 2010. Open issues in genetic  
1071 programming, *Genet. Program. Evol. M.*, 11, 339–363.
- 1072 Oehler, F., Coco, G., Green, M.O., Bryan, K.R., 2011. A data-driven approach to predict  
1073 suspended-sediment reference concentration under non-breaking waves. *Continental*  
1074 *Shelf Research* 46, 96-106.
- 1075 Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying  
1076 variable importance in artificial neural networks using simulated data. *Ecological*  
1077 *Modelling* 178, 389-397.
- 1078 Olden, J. D., Lawler, J. J., Poff, N. L., 2008. Machine learning methods without tears: a primer  
1079 for ecologists. *The Quarterly review of biology*, 83(2), 171-193.
- 1080 Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A., Moore, J. H., 2017. Data-driven advice for  
1081 applying machine learning to bioinformatics problems. *arXiv preprint arXiv:1708.05070*.
- 1082 Packard, N. H., Crutchfield, J. P., Farmer, J. D., Shaw, R. S., 1980. Geometry from a time series.  
1083 *Physical review letters*, 45(9), 712.
- 1084 Palmsten, M. L., Splinter, K. D., Plant, N. G., Stockdon, H. F., 2014. Probabilistic estimation of  
1085 dune retreat on the Gold Coast, Australia. *Shore and Beach*, 82(4), 35-43.
- 1086 Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Dierolf, U.,  
1087 2013. Making research data repositories visible: The re3data. org registry. *PloS*  
1088 *one*, 8(11), e78080.
- 1089 Pape, L., Ruessink, B. G., Wiering, M. A., Turner, I. L., 2007. Recurrent neural network  
1090 modeling of nearshore sandbar behavior, *Neural Networks*, 20, 509–518.
- 1091 Pape, L., Kuriyama, Y., Ruessink, B. G., 2010. Models and scales for cross-shore sandbar  
1092 migration, *J. Geophys. Res.*, 115, F03043, doi:10.1029/2009JF001644.
- 1093 Passarella, M., E. B. Goldstein, S. De Muro, G. Coco, 2018. The use of genetic programming to  
1094 develop a predictor of swash excursion on sandy beaches. *Nat. Hazards Earth Syst. Sci.*,  
1095 18, 599-611, <https://doi.org/10.5194/nhess-18-599-2018>
- 1096 Passeri, D. L., Hagen, S. C., Plant, N. G., Bilskie, M. V., Medeiros, S. C., Alizad, K., 2016. Tidal  
1097 hydrodynamics under future sea level rise and coastal morphology in the Northern Gulf

- 1098 of Mexico, *Earth's Future*, 4, 159–176, doi:10.1002/2015EF000332.
- 1099 Pfeiffenberger, H., Carlson, D., 2011. "Earth System Science Data" (ESSD)-A Peer Reviewed  
1100 Journal for Publication of Data. *D-Lib Magazine*, 17(1/2).
- 1101 Pinsky, M. L., Guannel, G., Arkema, K. K., 2013. Quantifying wave attenuation to inform  
1102 coastal habitat conservation, *Ecosphere*, 4(8).
- 1103 Plant, N. G., Holland, K. T., 2011. Prediction and assimilation of surf-zone processes using a  
1104 Bayesian network. Part I: Forward models, *Coastal Engineering*, 58(1), 119-130.
- 1105 Plant, N. G., Stockdon, H.F., 2012. Probabilistic prediction of barrier-island response to  
1106 hurricanes, *J. Geophys. Res.*, 117, F03015, doi:10.1029/2011JF002326.
- 1107 Plant, N. G., Aarninkhof, S. G. J., Turner, I. L., Kingston, K., 2007. The performance of  
1108 shoreline detection models applied to video imagery, *J. Coast. Res.*, 23(3), 658-670.
- 1109 Plant, N. G., Holman, R.A., Freilich, M.H., Birkemeier, W.A., 1999. A simple model for  
1110 interannual sandbar behavior, *Journal of Geophysical Research: Oceans*, 104(C7), 15755-  
1111 15776.
- 1112 Plant, N. G., Holland, K. T., Holman, R.A., 2006. A dynamical attractor governs beach response  
1113 to storms, *Geophysical Research Letters*, 33(17).
- 1114 Plant, N. G., Thieler, E. R., Passeri, D.L., 2016. Coupling centennial-scale shoreline change to  
1115 sea-level rise and coastal morphology in the Gulf of Mexico using a Bayesian network,  
1116 *Earth's Future*, 4(doi:10.1002/2015EF000331).
- 1117 Plant, N. G., Flocks, J., Stockdon, H.F., Long, J.W., Guy, K., Thompson, D.M., Cormier, J.M.,  
1118 Smith, C.G., Miselis, J.L., Dalyander, P. S., 2014. Predictions of barrier island berm  
1119 evolution in a time-varying storm climatology, *J. Geophys. Res. Earth Surf.*, 119, 300–  
1120 316, doi:10.1002/ 2013JF002871.
- 1121 Plomaritis, T. A., Costas, S., Ferreira, Ó., 2017. Use of a Bayesian Network for coastal hazards,  
1122 impact and disaster risk reduction assessment at a coastal barrier (Ria Formosa,  
1123 Portugal). *Coastal Engineering*.
- 1124 Poelhekke, L., Jäger, W. S., van Dongeren, A., Plomaritis, T. A., McCall, R., Ferreira, Ó., 2016.  
1125 Predicting coastal hazards for sandy coasts with a Bayesian network. *Coastal*  
1126 *Engineering*, 118, 21-34.
- 1127 Poli, R., Langdon, W. B., McPhee, N. F., 2008. A field guide to ge- netic programming, Lulu  
1128 Enterprises Uk Limited.

- 1129 Popper, K., Miller, D., 1983. A proof of the impossibility of inductive probability. *Nature*, 302,  
1130 687-688.
- 1131 Priestley, M. B., 1981. Spectral analysis and time series, 890 pp., Academic Press, London.
- 1132 Rigos, A., Tsekouras, G. E., Vousedoukas, M. I., Chatzipavlis, A., Velegrakis, A. F., 2016a. A  
1133 Chebyshev polynomial radial basis function neural network for automated shoreline  
1134 extraction from coastal imagery. *Integrated Computer-Aided Engineering*, 23(2), 141-  
1135 160.
- 1136 Rigos, A., Tsekouras, G. E., Chatzipavlis, A., Velegrakis, A. F., 2016b. Modeling Beach  
1137 Rotation Using a Novel Legendre Polynomial Feedforward Neural Network Trained by  
1138 Nonlinear Constrained Optimization. In IFIP International Conference on Artificial  
1139 Intelligence Applications and Innovations (pp. 167-179). Springer International  
1140 Publishing.
- 1141 Roelvink, D., Reniers, A., Van Dongeren, A. P., de Vries, J. V. T., McCall, R., Lescinski, J.,  
1142 2009. Modelling storm impacts on beaches, dunes and barrier islands. *Coastal*  
1143 *Engineering*, 56(11), 1133-1152.
- 1144 Ruessink, B. G., 2005. Calibration of nearshore process models: Application of a hybrid genetic  
1145 algorithm, *J. Hydroinformatics*, 7, 135–149.
- 1146 Schmidt, M., Lipson, H., 2009. Distilling free-form natural laws from experimental data,  
1147 *Science*, 324, 81–85.
- 1148 *Scientific Data*, 2014. More bang for your byte. *Sci. Data*, 1, 140010.
- 1149 Sherman, D. J., 1995. Problems of scale in the modeling and interpretation of coastal  
1150 dunes. *Marine Geology*, 124(1-4), 339-349.
- 1151 Splinter, K. D., Holman, R. A., Plant, N.G., 2011. A behavior-oriented dynamic model for  
1152 sandbar migration and 2DH evolution, *Journal of Geophysical Research C: Oceans*,  
1153 116(1).
- 1154 Splinter, K. D., Turner, I. L., Davidson, M. A., 2013. How much data is enough? The importance  
1155 of morphological sampling interval and duration for calibration of empirical shoreline  
1156 models. *Coastal Engineering*, 77, 14-27.
- 1157 Stalzer, M., Mentzel, C., 2016. A preliminary review of influential works in data-driven  
1158 discovery. *SpringerPlus*, 5(1), 1266.



- 1159 Stephens, S. A., Coco, G., Bryan, K. R., 2011. Numerical Simulations of Wave Setup over  
1160 Barred Beach Profiles: Implications for Predictability, *Journal of Waterway, Port,*  
1161 *Coastal and Ocean Engineering*, 137(4), 175-181.
- 1162 Stockdon, H. F., Thompson, D. M., Plant, N. G., Long, J. W., 2014. Evaluation of wave runup  
1163 predictions from numerical and parametric models, *Coastal Engineering*, 92, 1-11.
- 1164 Sugihara, G., May, R. M., 1990. Nonlinear forecasting as a way of distinguishing chaos from.  
1165 *Nature*, 344, 6268.
- 1166 Sugihara, G., May, R., Ye, H., Hsieh, C. H., Deyle, E., Fogarty, M., Munch, S., 2012. Detecting  
1167 causality in complex ecosystems. *science*, 338(6106), 496-500.
- 1168 Takens, F., 1981. Detecting strange attractors in turbulence. *Lecture notes in mathematics*,  
1169 898(1), 366-381.
- 1170 Teodoro, A.C, Veloso-Gomes, F. and Goncalves H, 2007. Retrieving TSM Concentration From  
1171 Multispectral Satellite Data by Multiple Regression and Artificial Neural Networks,  
1172 *IEEE Transactions on Geoscience and Remote Sensing*, 45, 5.
- 1173 Tinoco, R. O., Goldstein, E. B., Coco, G., 2015. A data-driven approach to develop physically  
1174 sound predictors: Application to depth-averaged velocities on flows through submerged  
1175 arrays of rigid cylinders. *Water Resources Research*, 51(2), 1247-1263.
- 1176 Tsekouras, G. E., Rigos, A., Chatzipavlis, A., Velegrakis, A., 2015. A neural-fuzzy network  
1177 based on Hermite polynomials to predict the coastal erosion. In *Engineering Applications*  
1178 *of Neural Networks* (pp. 195-205). Springer International Publishing.
- 1179 Turner, I. L., Harley, M. D., Short, A. D., Simmons, J. A., Bracs, M. A., Phillips, M. S., Splinter,  
1180 K. D., 2016. A multi-decade dataset of monthly beach profile surveys and inshore wave  
1181 forcing at Narrabeen, Australia. *Scientific Data*, 3.
- 1182 Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental  
1183 modelling. *Ecological modelling*, 203(3), 312-318.
- 1184 Valentine, A. P., Kalnins, L. M., 2016. An introduction to learning algorithms and potential  
1185 applications in geomorphometry and earth surface dynamics. *Earth surface dynamics.*, 4,  
1186 445-460.
- 1187 van der Werf, J. J., Schretlen, J. J., Ribberink, J. S., O'Donoghue, T., 2009. Database of full-scale  
1188 laboratory experiments on wave-driven sand transport processes. *Coastal*

- 1189            *Engineering*, 56(7), 726-732.
- 1190 van Maanen, B., Coco, G., Bryan, K.R., Ruessink, B.G., 2010. The use of artificial neural  
1191            networks to analyze and predict alongshore sediment transport. *Nonlinear Processes in*  
1192            *Geophysics*: 17, p. 395-404. doi:10.5194/npg-17-395-2010
- 1193 Van Verseveld, H. C. W., Van Dongeren, A. R., Plant, N. G., Jäger, W. S., den Heijer, C., 2015.  
1194            Modelling multi-hazard hurricane damages on an urbanized coast with a Bayesian  
1195            Network approach. *Coastal Engineering*, 103, 1-14.
- 1196 Werner, B. T., 1999. Complexity in natural landform patterns. *Science*, 284(5411), 102-104.
- 1197 Wikle, C. K., Berliner, L. M., 2007. A Bayesian tutorial for data assimilation, *Physica D*, 230, 1-  
1198            16.
- 1199 Wilson, K. E., Adams, P. N., Hapke, C. J., Lentz, E. E., Brenner, O., 2015. Application of  
1200            Bayesian Networks to hindcast barrier island morphodynamics. *Coastal*  
1201            *Engineering*, 102, 30-43.
- 1202 Winant, C. D., Inman, D.L., Nordstrom, C. E., 1975. Description of seasonal beach changes  
1203            using empirical eigenfunctions, *Journal of Geophysical Research*, 80(15), 1979-1986.
- 1204 Wright, L. D., Short, A. D., 1984. Morphodynamic variability of surf zones and beaches: a  
1205            synthesis. *Marine Geology*, 56(1-4), 93-118.
- 1206 Wright, L. D., Short, A. D., Green, M. O., 1985. Short-term changes in the morphodynamic  
1207            states of beaches and surf zones: An empirical predictive model, *Mar. Geol.*, 62, 339-  
1208            364.
- 1209 Yan, B., Zhang, Q., Wai, O.W.H., 2008. Prediction of sand ripple geometry under waves using  
1210            an artificial neural network. *Computers & Geoscience* 34, 1655-1664.
- 1211 Yao, X., 1999. Evolving artificial neural networks. *Proceedings of the IEEE*, 87(9), 1423-1447.
- 1212 Yates, M.L., Le Cozannet, G., 2012. Evaluating European coastal evolution using Bayesian  
1213            networks. *Natural Hazards and Earth Systems Sciences* 12, 1173–1177
- 1214 Yates, M. L., Guza, R.T., O'Reilly, W. C., 2009. Equilibrium shoreline response: Observations  
1215            and modeling, *Journal of Geophysical Research C: Oceans*, 114(9).
- 1216 Ye, H., Sugihara, G., 2016. Information leverage in interconnected ecosystems: Overcoming the  
1217            curse of dimensionality. *Science*, 353(6302), 922-925.
- 1218 Yoon, H-D, Cox, D.T., Kim, M., 2013. Prediction of time-dependent sediment suspension in the  
1219            surf zone using artificial neural network. *Coastal Engineering*, 71, 78–86,

**\*\*\*Please note that this is a non-peer reviewed EarthArXiv preprint\*\*\***

- 1220 <http://dx.doi.org/10.1016/j.coastaleng.2012.08.005>
- 1221 Zeigler, S. L., Thieler, E.R., Gutierrez, B.T., Plant, N.G., Hines, M., Fraser, J.D., Catlin, D.H.,
- 1222 Karpanty, S.M., 2017. Smartphone Technologies and Bayesian Networks to Assess
- 1223 Shorebird Habitat Selection, *Wildlife Society Bulletin*; DOI: 10.1002/wsb.820.
- 1224 Zdeborová, L., 2017. Machine learning: New tool in the box. *Nature Physics*, 13(5), 420-421.