Earth Surf. Dynam., 2, 67–82, 2014
www.earth-surf-dynam.net/2/67/2014/
doi:10.5194/esurf-2-67-2014

Earth Surface Dynamics
Open Access

Data-driven components in a model of inner-shelf sorted bedforms: a new hybrid model

E. B. Goldstein[1], G. Coco[2], A. B. Murray[1], and M. O. Green[3]

[1]Division of Earth and Ocean Sciences, Nicholas School of the Environment, Center for Nonlinear and Complex Systems, Duke University, P.O. Box 90227, Durham, NC 27708, USA
[2]Environmental Hydraulics Institute, "IH Cantabria", c/Isabel Torres no. 15, Universidad de Cantabria, 39011 Santander, Spain
[3]National Institute of Water and Atmospheric Research (NIWA), P.O. Box 11-115, Hamilton, New Zealand

Correspondence to: E. B. Goldstein (evan.goldstein@duke.edu)

Received: 28 September 2013 – Published in Earth Surf. Dynam. Discuss.: 23 October 2013
Revised: 13 January 2014 – Accepted: 16 January 2014 – Published: 28 January 2014

Abstract. Numerical models rely on the parameterization of processes that often lack a deterministic description. In this contribution we demonstrate the applicability of using machine learning, a class of optimization tools from the discipline of computer science, to develop parameterizations when extensive data sets exist. We develop a new predictor for near-bed suspended sediment reference concentration under unbroken waves using genetic programming, a machine learning technique. We demonstrate that this newly developed parameterization performs as well or better than existing empirical predictors, depending on the chosen error metric. We add this new predictor into an established model for inner-shelf sorted bedforms. Additionally we incorporate a previously reported machine-learning-derived predictor for oscillatory flow ripples into the sorted bedform model. This new "hybrid" sorted bedform model, whereby machine learning components are integrated into a numerical model, demonstrates a method of incorporating observational data (filtered through a machine learning algorithm) directly into a numerical model. Results suggest that the new hybrid model is able to capture dynamics previously absent from the model – specifically, two observed pattern modes of sorted bedforms. Lastly we discuss the challenge of integrating data-driven components into morphodynamic models and the future of hybrid modeling.

1 Introduction

Parameterizations become necessary in morphodynamic models when processes cannot be described entirely from conservation laws. This is often the case with descriptions of sediment transport, where the mechanics are multidimensional and highly nonlinear (e.g., have thresholds). Parameterizations are often developed through the collection and processing of experimental data. This results in formulas that, because they have been developed through inductive methods, are subject to many caveats: constraints regarding the applicable forcing conditions or the appropriate setting for use. The inaccuracy of individual predictors has significant consequences in nonlinear morphodynamic models because errors accumulate as inaccuracy is (1) propagated through the nonlinear pieces of the model (e.g., Bolaños et al., 2012) and (2) propagated in time (e.g., Pape et al., 2010).

Some prediction schemes may perform well only in specific settings or under specific hydrodynamic conditions (Cacchione et al., 2008; Bolaños et al., 2012). This is an example of locally optimal predictors, performing well with a single set of data but not necessarily transferable to other settings (both physical locations and hydrodynamic conditions). The existence of many locally optimal predictors (each developed from its own data set) leads to the problem of selecting the appropriate predictor for a morphodynamic model. One solution to this difficulty is to sidestep it entirely and instead develop globally optimal predictors from multi-setting data sets that encompass wide ranges of forcing conditions and independent variables. The hope is that differences in locally optimal solutions may be attributed to an independent variable that may become apparent when building a single, unified globally optimal model.

The construction of globally optimal predictors is difficult because large multi-setting data sets with nonlinear relationships and multiple independent variables are difficult to visualize and interpret. Traditional techniques for developing successful parameterizations include converting multidimensional data sets into low-dimensional spaces and then fitting a curve. However, collapsing data into combined parameters may inherently bias the resultant predictor and may obscure subtle relationships in the data. One method to detect relationships in large, nonlinear, multidimensional data sets is machine learning (ML), a class of computational optimization routines. A range of ML techniques have previously been used successfully to develop data-driven parameterizations: artificial neural networks (ANN) have been used to parameterize alongshore suspended sediment transport in the surf zone (van Maanen et al., 2010), sediment suspension in the surf zone (Yoon et al., 2013), and near-bed reference concentration (Oehler et al., 2012). Boosted regression trees (BRT) have been used to parameterize suspended sediment reference concentration (Oehler et al., 2012), and genetic programming techniques have been used to develop predictions of wave-generated ripple geometry (Goldstein et al., 2013), roughness in vegetated flows (Baptist et al., 2007), and fluvial sediment transport (Kitsikoudis et al., 2013). Aside from small-scale process descriptions, data-driven approaches have also been used as stand-alone morphodynamic models (Pape et al., 2007, 2010) and to calibrate model parameters (Knaapen and Hulscher, 2002, 2003; Ruessink, 2005).

In this contribution we focus on the data-driven prediction of near-bed reference concentration under unbroken waves. As the bottom boundary condition for calculating suspended sediment transport, reducing error is of paramount importance for accurate predictions of total suspended sediment load. Several parameterizations already exist, notably Nielsen (1986) and Lee et al. (2004). Recent work by Oehler et al. (2012) demonstrated the ability of ML predictors to outperform traditional empirical prediction schemes for reference concentration (i.e., Lee et al., 2004; Nielsen, 1986). The BRT and ANN model developed by Oehler et al. (2012) is an accurate predictor of reference concentration, but the predictor is not smooth, physically interpretable, or economical in length; all problems when attempting to incorporate the results into a morphodynamic model. Here we use genetic programming (GP) to develop a smooth and physically interpretable parameterization of near-bed reference concentration. GP is a population-based optimization technique where the population is composed of individual predictors (Koza, 1992). Using evolutionary principles (e.g., crossover, mutation) to develop new solutions, the functional form of the predictor and the location and presence of the variables within a given predictor are adjusted and optimized to find a globally optimum solution.

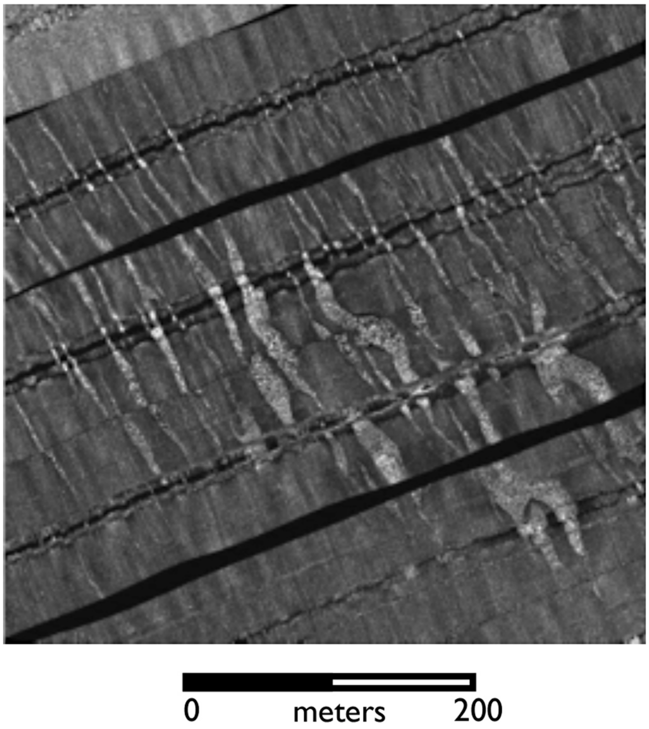

Figure 1. Sorted bedforms present in ~ 5 m of water off the coast of Tairua Beach, New Zealand (Coco et al., 2007a). White areas are composed of coarse sediment, while dark areas are floored by fine sediment. Shoreline is towards the bottom of the panel.

The development of a new near-bed suspended sediment reference concentration predictor using GP is the first objective of this work. The second objective is to incorporate this new predictor (and a previously developed predictor for ripple geometry, built with GP) into a previously developed model of inner-shelf sorted bedforms (Coco et al., 2007a) to develop a “hybrid” numerical model (Krasnopolsky and Fox-Rabinovitz, 2006), where data-driven components are combined with widely accepted formulas for hydrodynamics and sediment transport. Previous examples of the hybrid approach are found in studies of shoreline change (Karunarathna and Reeve, 2013), hydrology (Corzo et al., 2009), and the atmospheric and climate system (Krasnopolsky and Fox-Rabinovitz, 2006).

Spatially extensive (kilometer scale) patches of segregated coarse and fine-grained sediment (Fig. 1) with only slight bathymetric relief (centimeter to meter scale) relative to bedform pattern wavelength (10 m–km) are present on many continental shelf systems (Coco et al., 2007b). Unlike most bedforms that develop solely as an interaction between bathymetry and flow, recent work implicates a sorting feedback as the mechanism for the development of inner-shelf

Table 1. Summary of experiments used in this study.

Study	Mean water depth (m)	Sediment grain size (mm)	Sampling rate (Hz)	Burst duration (min)	Sorted bedform field?
(Green, 1999; Green and Black, 1999)	7	0.23	4–5	10–17.06	No
(Green et al., 2004; Trembanis et al., 2004)	15	0.22	1	15	Yes
(Green et al., 2004; Trembanis et al., 2004)	22	0.22	4	8.5	Yes
(Green et al., 2004; Trembanis et al., 2004)	22	0.75	1	15	Yes
(Vincent and Green, 1999)	25	0.33	4	10	No
(Green and MacDonald, 2001)	1.7	0.15	4–5	4.267–5	No

"sorted bedforms" (Murray and Thieler, 2004; Coco et al., 2007a, b). The sorting feedback is initiated by wave-generated ripples whose size is a function of seabed composition and hydrodynamic forcing conditions (e.g., Cummings et al., 2009). Regions covered with fine sediment support smaller wave-generated ripples than areas mantled by coarse sediment. Strong turbulence above the large wave ripples on coarse domains enhances the erosion of fine material from the bed (and also functions as a barrier to the deposition of suspended fine sediment). Near-bottom currents lead to the advection of suspended fine material and the preferential settling of suspended fine sediment in areas where the seabed is composed of predominantly fine sediment with small wave ripples (and correspondingly less turbulence induced by the smaller features). Through self-organization this local sorting feedback leads to spatially extensive features. The numerical model of Coco et al. (2007a) indicates that the sorting feedback operates in a wide range of forcing conditions (Coco et al., 2007b).

Sorted bedforms show several configurations that we divide into two distinct end-member patterns typified by the location of the coarse domain, either in the trough of the bedform or on the flanks of the bedforms (appearance on both the updrift and/or downdrift are possible; e.g., Goff et al., 2005; Ferrini and Flood, 2005). We note that within an individual sorted bedform field the pattern configuration can change (Thieler et al., 2014; Ferrini and Flood, 2005). Previous work with the finite-amplitude models by Murray and Thieler (2004) and Coco et al. (2007a) showed the presence of coarse domains solely on the downdrift flank of bedforms. While Coco et al. (2007b) did show the potential for coarse domains to occur in the trough of bedforms, this configuration was highly path dependent (i.e., the result of a high wave event that is preceding and followed by smaller waves). Van Oyen et al. (2010, 2011), through linear stability analysis, showed the presence of two pattern modes in the initial infinitesimal-amplitude instability that correspond to these two distinct configurations. However Van Oyen et al. (2010, 2011) showed that each pattern mode is the result of separate feedback mechanisms, where coarse domains present in troughs occurred as the result of a flow–bathymetry feedback, while coarse domains present on bedform flanks is the result of the previously described sorting feedback (refereed to as the "roughness" feedback by Van Oyen et al. (2010, 2011).

With the goal of presenting a new hybrid model, we first describe the development of the near-bed suspended sediment reference concentration predictor from the large data set of Green and colleagues (Green, 1996, 1999; Green and Black, 1999; Vincent and Green, 1999; Green and MacDonald, 2001; Green et al., 2004; Trembanis et al., 2004). We then outline the sorted bedform model and the modifications to incorporate the new data-driven components. This new model is meant as an update to the Coco et al. (2007a) model. The new predictors in the hybrid model are more accurate and better performing than the formulations used in the Coco et al. (2007a) model. Finally, we present a novel experiment with the new hybrid model to show autogenic behaviors that were not present in the Coco et al. (2007a) model (i.e., the appearance of two pattern configurations solely from a sorting feedback) and discuss advantages and disadvantages of this data-driven approach.This paper does not attempt to quantitatively compare the new hybrid model against older modeling efforts: instead we offer this new model as a refinement to the previous model that is additionally able to capture new dynamics.

2 GP methods

2.1 Data set

Figure 2 shows the multi-setting field data set composed of 1748 individual measurements from 6 separate field experiments at different locations in New Zealand. We briefly summarize the experiments below and in Table 1; a detailed summary of each experiment and the specific methodology used to determine the near-bed suspended sediment reference concentration (C_0; g L^{-1}), significant near-bed orbital velocity (U_{sig}; m s^{-1}), wave orbital diameter at the bed (d_0; m), mean grain size (d_{50}; m), and mean spectral wave period at the bed (T_{mean}; s) is available in the associated references. A single experiment (Green and Black, 1999; Green, 1999) collected 127 measurements seaward of the surf zone with mean water

Table 2. Solutions for reference concentration.

Solution	Complexity	MSE
$C_0 = 0.182$	1	0.070
$C_0 = U_{\mathrm{sig}}^2$	2	0.057
$C_0 = 0.637 U_{\mathrm{sig}}$	3	0.056
$C_0 = \left(1.19 U_{\mathrm{sig}}\right)^2$	4	0.052
$C_0 = U_{\mathrm{sig}} - 0.647\,(1000 d_{50})$	5	0.048
$C_0 = \left(\frac{0.235 U_{\mathrm{sig}}}{\left(1000 d_{50}\right)}\right)^2$	7	0.048
$C_0 = \left(\frac{0.328 U_{\mathrm{sig}}}{0.0688+\left(1000 d_{50}\right)}\right)^2$	9	0.045
$C_0 = \left(1.27\sqrt{U_{\mathrm{sig}}} - 1.21\,(1000 d_{50})\right)^2$	12	0.045
$C_0 = \frac{0.179 U_{\mathrm{sig}}^2 - 0.00538}{d_0\left(1000 d_{50}\right)} + \frac{0.0185+0.179 U_{\mathrm{sig}}^2 d_0 - 0.179 U_{\mathrm{sig}}^2 - 0.0319 U_{\mathrm{sig}}^4}{\left(1000 d_{50}\right)}$	41	0.043

depth of 7 m. Data from three experiments (Green et al., 2004; Trembanis et al., 2004) were collected from separate locations in a field of sorted bedforms (669, 126, and 554 measurements). A single instrument frame was located in a domain composed of coarse sand (22 m depth) and two instrument frames were located in fine sand domains (15 and 22 m depth). The fifth experiment was deployed off of a headland in 25 m of water depth (56 measurements; Vincent and Green, 1999). The final experiment in the database collected 241 measurements in a microtidal estuary in a mean water depth of 1.7 m (Green and MacDonald, 2001). All data were gathered in burst mode, with burst durations ranging from 4.267 to 17.06 min. In addition to the multiple settings and significant amount of data, this data set is ideal for application in the sorted bedform model because three of the six experiments in the composite data set are derived from a sorted bedform field (Green et al., 2004; Trembanis et al., 2004).

2.2 Selection of training, validation, and testing data sets

The database is split into three subsets to be used as training, validation, and testing. The training data set is used to develop candidate solutions. The validation data set is used to evaluate the generality of a predictor, the fitness of GP-derived solutions against more data, and ultimately to determine which predictors persist. The testing data set is unused and unseen by the GP algorithm; it is reserved as an independent test of the final predictors (and other published predictors). Because our database does not cover the entirety of the forcing space with equal density (Fig. 3), the selection and partitioning of data into these three categories is crucial for developing a well-performing predictor applicable to a range of environments (e.g., Bowden et al., 2002). The C_0 data set is sparse in areas because of a lack of collected data, while dense in other regions of phase space as a result of similar field settings, forcing conditions, and the number of data points collected in a given experiment. If the data are randomly divided, there is a potential that the training data exclude data from sparse regions in the data set (i.e., coarse-grained and/or strong hydrodynamic data). However, in the genetic programming literature we could find no proven “best practice” for selection of the data subsets or an optimal percentage of training, validation, and testing data (Kuschu, 2002; Panait and Luke, 2003; Gagné et al., 2006); we therefore use a technique that was successful in a previous study (Goldstein et al., 2013).

Informed data selection has been shown to produce better results with ML predictors than “blind” or random data selection (e.g., Bowden et al., 2002; May et al., 2010). In this study we select training data through the use of a maximum dissimilarity algorithm (MDA; Camus et al., 2011). This algorithm is not a clustering routine (where centroids denote a representative value of the data in the cluster), but is instead a selection routine (where a centroid represents the most dissimilar data point from the previous centroids; Camus et al., 2011). This selection routine allows the use of a minimum of training data that is able to capture the variance present in the entire data set while leaving the majority of the data to be utilized as validation and testing.

The maximum dissimilarity algorithm is described in Camus et al. (2011) and we review the method. Selection starts with the linear normalization of the independent variables to a value between 0 (minimum value of a given variable) and 1 (maximum value of a given variable). A single data point, a “seed”, is selected as the first centroid. The algorithm then selects the additional centroids (the number determined by the user) through an iterative process: each data point is a four-dimensional vector (normalized T_{mean}, U_{sig}, d_0, d_{50} space) and is associated with a distance to the nearest centroid. The single data point with the maximum distance between itself and the nearest centroid is selected as the next centroid (Camus et al., 2011). The MDA routine continues until the user-defined number of centroids is reached and the data are then denormalized.

There remains significant ambiguity in determining the appropriate number of centroids needed to accurately represent

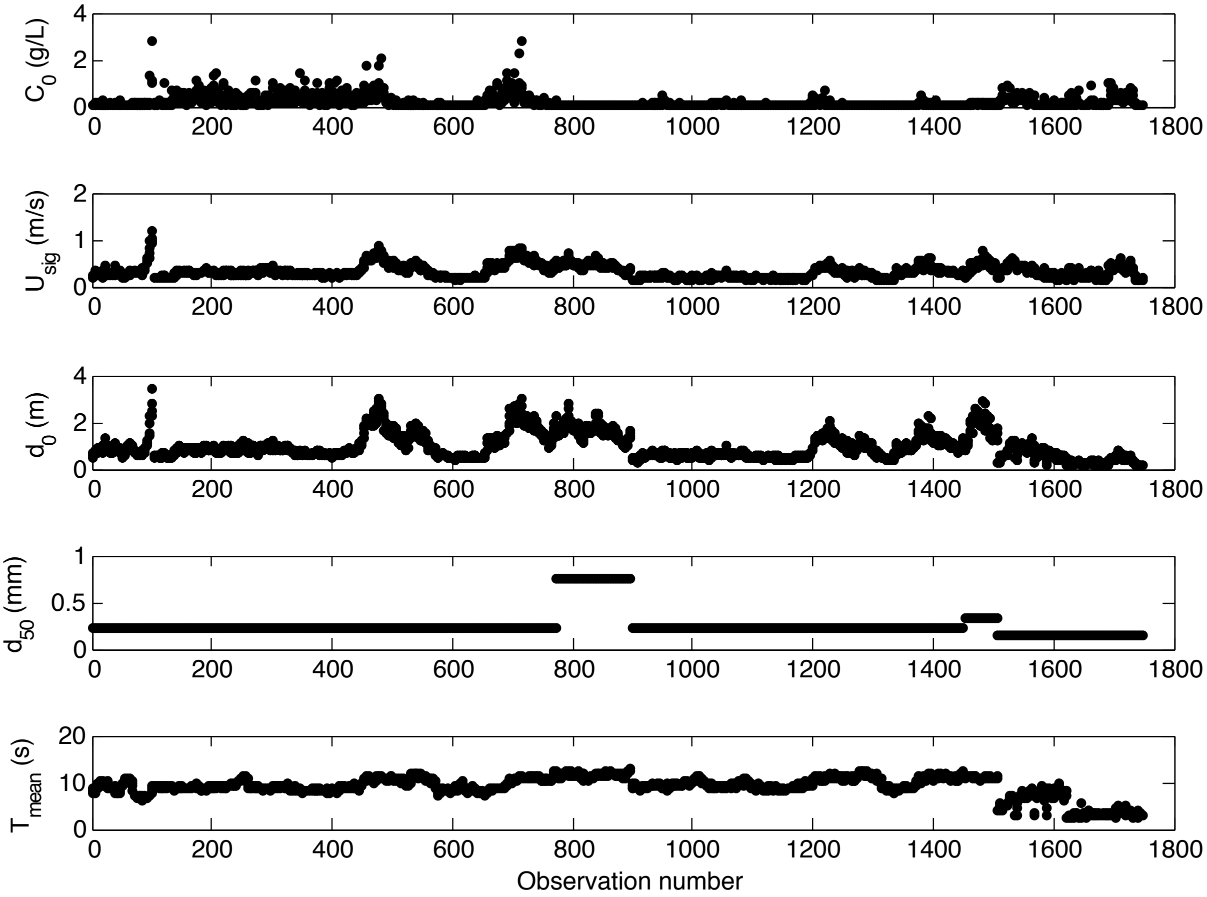

Figure 2. Observations of suspended sediment reference concentration data set C_0 and concomitant measurements of significant wave velocity at the bed (U_{sig}), wave orbital excursion at the bed (d_0), mean grain size of bed material (d_{50}), and mean spectral wave period at the bed (T_{mean}). Note that mean grain size of bed material is shown here in millimeters. A similar figure appears in Oehler et al. (2012).

a data set, especially continuous data (e.g., May et al., 2010; Goldstein et al., 2013). Selecting too many centroids can rob the validation and testing data sets of poorly represented data (e.g., large T_{mean}, U_{sig}, d_0, d_{50}) and may tend to cause the GP to produce overly complex predictors (e.g., Gonçalves and Silva, 2013; Oates and Jensen, 1997, 1998). The selection of too few centroids can leave the testing data with too few data points to capture the variability in the data set (Goldstein et al., 2013). We use 40 centroids for the prediction of C_0 (centroid locations can be seen in Fig. 3), the same as Goldstein et al. (2013). Data selected as the centroid locations are used for the training data, while the remaining data are used for validation and testing data. The data set is split between validation and testing randomly, without using a selection routine. The final breakdown for the data sets is ~ 2 % training, ~ 49 % validation, and ~ 49 % testing.

2.3 Genetic programming

We operate on this data set using the ML technique of genetic programming (GP; Koza, 1992; Poli et al., 2008), where candidate solutions (i.e., randomly generated initial equations) are evaluated and subsequently modified by adjusting the independent variables as well as the mathematical relationships between variables (i.e., the mathematical form). Independent variables used in this study to predict C_0 are T_{mean}, U_{sig}, d_0, and d_{50}. We use T_{mean}, U_{sig}, and d_0 as separate independent variables for input to the GP (though they are related) in an attempt to introduce no additional information about which of these parameters is most relevant. Mathematical operators used in this study are + (addition), − (subtraction), × (multiplication), ÷ (division), and $\sqrt{}$ (square root), as well as integer powers (e.g., x^2, x^3, etc.). We omit logical functions in this analysis (e.g., if-then-else) because we aim to develop a smooth final solution.

Candidate solutions are evaluated based on a "fitness function", a user-defined error metric that determines how well a given candidate fits the validation data. Mean squared error (MSE) is used as the fitness function:

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n}(p_i - b_i)^2}{n}, \tag{1}$$

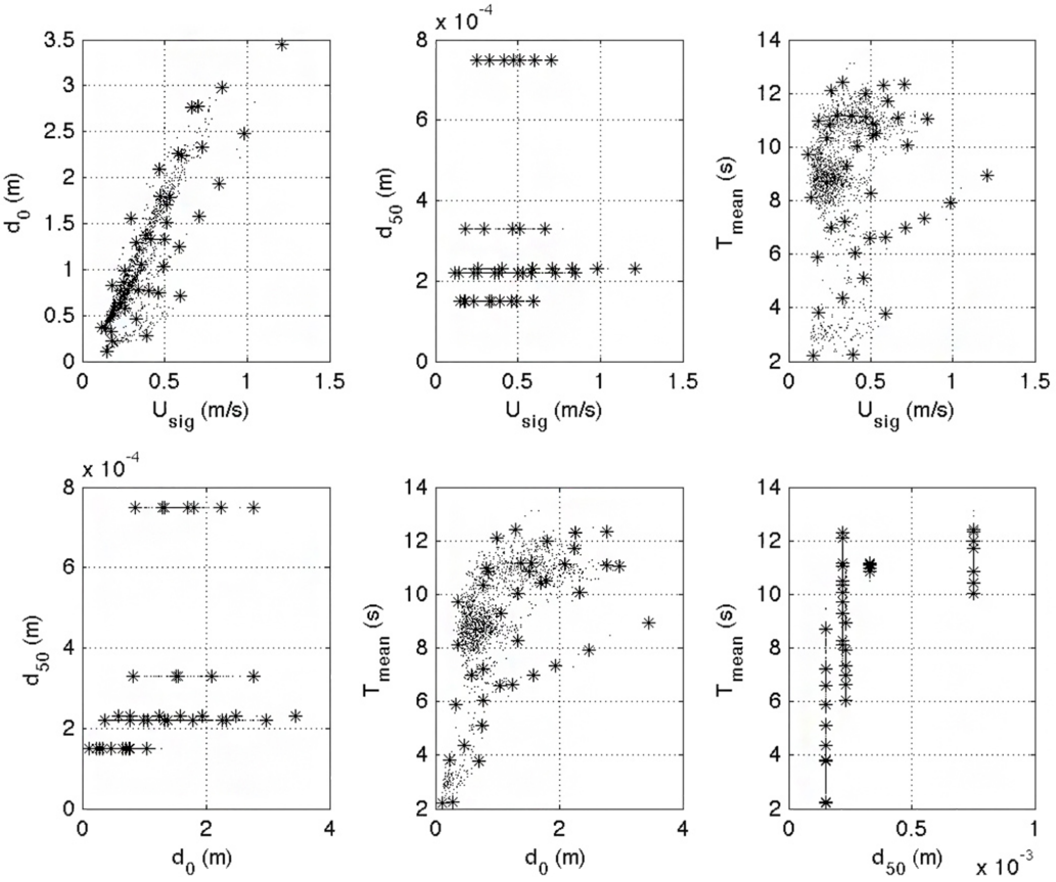

Figure 3. Visualization of the range of conditions in the C_0 data set. Each plot represents a two-dimensional projection of the entire data set onto the set of axes shown. For instance, the first panel with data projected onto the $d_0 - U_{\mathrm{sig}}$ plane shows no information about d_{50} or T_{mean}. Stars denote centroid locations (training data), while points denote unselected data (validation and testing). Note that centroids are distributed throughout the data set.

where n is the sample size, p are the predicted values, and b are the observed values. Candidate solutions that minimize mean squared error are retained and poor performing solutions are discarded. Retained solutions are rearranged, combined, and manipulated in a probabilistic manner according to combinatorial processes: solutions "crossover" by combining elements of other solutions to develop a new solution and "mutations" develop new mathematical expression to substitute or tack on to a previous solution. Candidate solutions are commonly encoded in GP software as graphs or "trees". The evolutionary processes that modify candidate solutions (change of variables and/or mathematical expression) is accomplished by adjusting tree "limbs" (Fig. 4). Predictors range from simple (small trees) to complex (large trees) as they are recombined in a variety of ways. The range of candidate solutions enables the searching of a large solution space, and the search process continues until a solution with zero error is found or the routine is halted.

In this study we use a proven software package developed by Schmidt and Lipson (2009, 2013). This software package, "Eureqa", outputs a suite of solutions with increasing mathematical "complexity", where complexity is a count of the numbers of operators and variables are used in the candidate solution. Each solution of a given complexity represents the equation with the least error compared to identically "complex" candidate solutions. Additionally, solutions must have less error compared to all previous less complex solutions. The line that traces the suite of solutions in complexity–fitness space is the "Pareto front", and is a graphical representation of increasing fitness with increasing complexity. Many predictors along the Pareto front, from simple to complex, are retained in the solution set, requiring the user to pick a single solution as the final predictor of choice.

In the results presented here there is no single zero-error solution found; therefore we cease the search after roughly 10^{10} formulas have been evaluated – continued search shows only marginal increases in predictive power (and this increase occurs only on more complex, likely overfitted, predictors). Several methods exist for eliminating overfitted solutions (e.g., Gonçalves et al., 2012). We use several

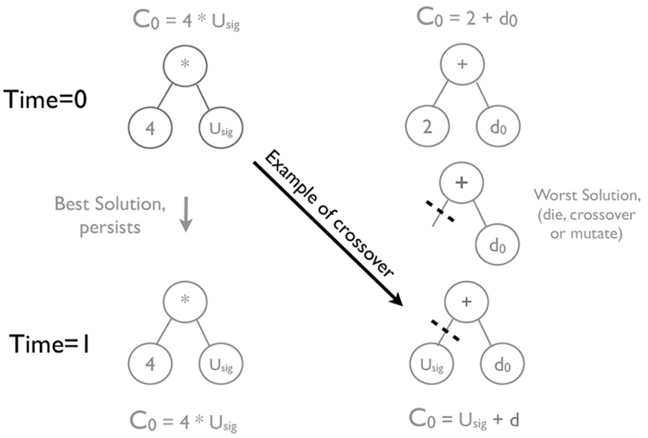

Figure 4. Example of the genetic programming process. Potential solutions are encoded as a population of trees. Here a hypothetical population of two solutions is shown. The first solution has a low MSE and therefore persists to the next iteration. The second solution has a high MSE and therefore is subject to removal, mutation, or crossover. An example of "crossover" is shown here, whereby the old solution is combined with parts of other, better performing solutions to create a new potential solution in the next iteration.

techniques in parallel to determine a single appropriate solution: (1) bias toward shorter, physically reasonable solutions, (2) examining "cliffs" in the Pareto front, and (3) examination of solution fit.

Compact, simple solutions tend to offer more generalization power and are likely less overfitted (the minimum description length principle; e.g., O'Neill et al., 2010). Additionally, shorter solutions reappear with repeat initialization of the genetic programming algorithm, suggesting that these reappearing candidates represent the globally optimum solutions for a given function size. Longer solutions do not tend to reappear, a result of a large search space that is not repeated during repeat initializations or the presence of multiple, equally optimal solutions in the large phase space (i.e., local minima). The inherent reproducibility of simple, weakly nonlinear solutions suggests their use as predictors until further data can be used to justify the use of highly nonlinear predictors.

Areas along the Pareto front where large gains in prediction are obtained with small gains in solution complexity, "cliffs", are a natural place to observe potential solutions (Fig. 5). Schmidt and Lipson (2009) observed many physically relevant solutions at the bottom of the last cliff of a given Pareto front, and therefore we focus our search for a final solution at the cliffs. Additionally, as candidate solutions are evaluated by minimizing error functions, solutions occasionally minimize mean squared error but are unphysical (e.g., functions that have poor extrapolation ability beyond the domain of the training data). These solutions must be manually disregarded, as there is as yet no means of excluding them.

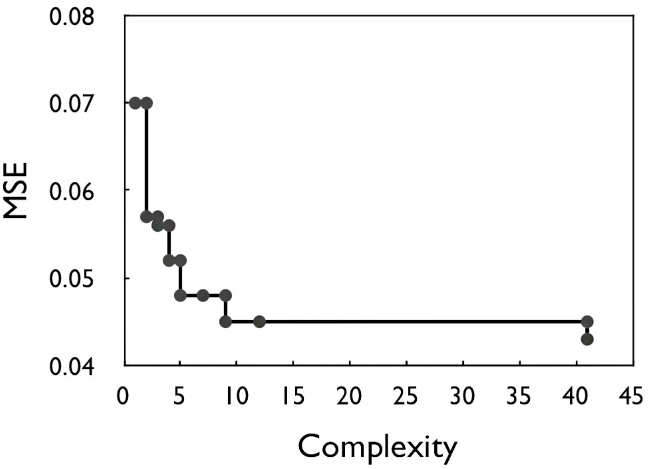

Figure 5. Reference concentration Pareto front; MSE is mean squared error of candidate solution versus the validation data set. Complexity is a quantification of the candidate solution length (both mathematical operators and variables).

Once a single predictor is selected, it is evaluated using the independent testing data (data that the ML algorithm has not seen) with the normalized root-mean-squared error (NRMSE):

$$\mathrm{NRMSE} = \frac{\sqrt{\mathrm{MSE}}}{\overline{b}}, \tag{2}$$

where $\overline{b}$ is the mean of the observed values. Additionally we report the correlation coefficient (Pearson's r) for each predictor evaluated against the independent testing data. The NRMSE and correlation coefficient are also reported for the reference concentration predictor of Nielsen (1986) and Lee et al. (2004) evaluated against the independent testing data.

3 GP results

The GP algorithm output is shown in Table 2 (note that numerical coefficients listed in the table are dimensional). This experiment evaluated 10^{10} formulas to develop the Pareto front shown in Fig. 5. Cliffs occur along the Pareto front at complexities of 2, 4, 5, 9, and 41 (Fig. 5). Predictors generally show nonlinear dependence on U_{sig}/d_{50}, qualitatively similar to the predictors developed by Nielsen (1986) and Lee et al. (2004), which both show dependence on the modified Shields parameter. We focus our analysis on the last cliff before the proliferation of very complex, nonlinear terms (solution 9):

$$C_0 = \left(\frac{0.328U_{\mathrm{sig}}}{0.0688 + (1000d_{50})}\right)^2. \tag{3}$$

Note that the coefficients of Eq. (3) are dimensional. Reserved testing data are used as an independent data set to

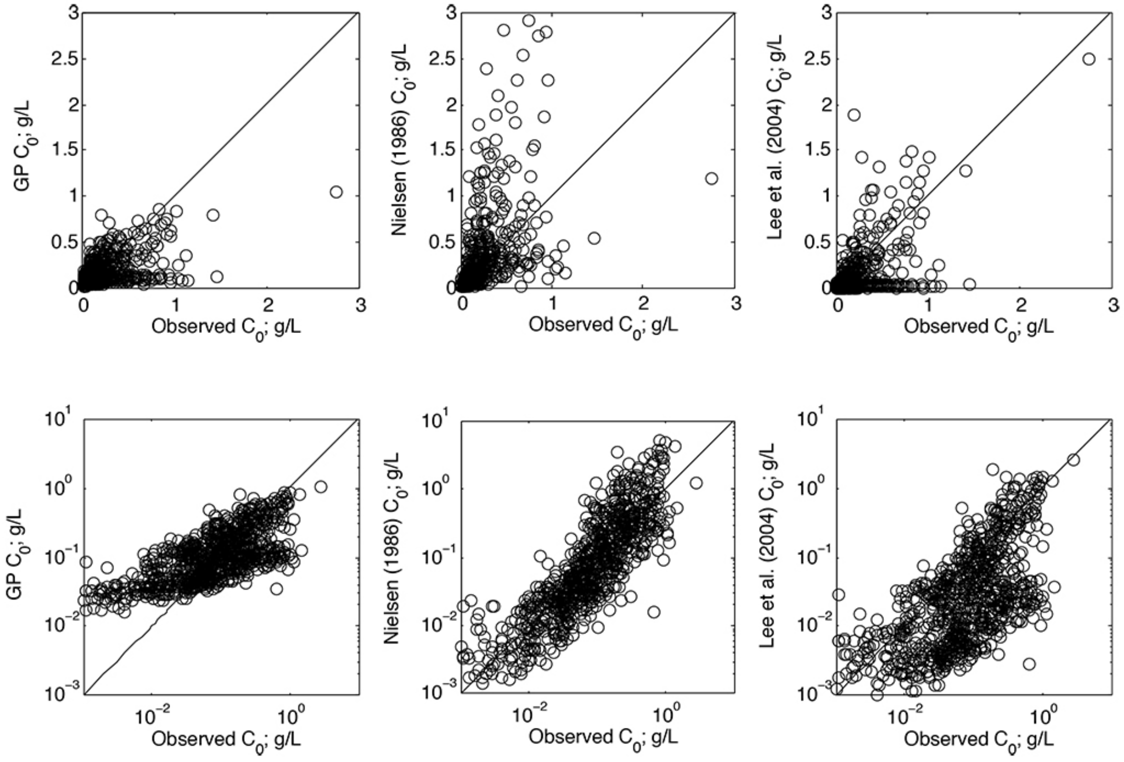

Figure 6. GP predictor of C_0, Nielsen (1986) and Lee et al. (2004) predictor evaluated using only the independent testing data set. Top row shows the predictors in linear space; bottom row shows log–log space.

compare the GP predictor as well as those developed by Nielsen (1986) and Lee et al. (2004): the NRMSE for each predictor is 1.1, 2.6, and 1.3, respectively, and the correlation coefficient is 0.58, 0.58, and 0.57, respectively. Results are shown in Fig. 6. The GP-derived predictor outperforms other predictors based on the NRMSE and is roughly identical to the other predictors based on correlation coefficient. However, we note that at very low concentrations the performance of Eq. (3) deteriorates.

4 Hybrid sorted bedform model overview

We now incorporate this new C_0 predictor into a previously described model of inner-shelf sorted bedforms developed by Coco et al. (2007a) that is based on the initial work of Murray and Thieler (2004). We briefly review the model below; a detailed treatment of the sediment transport relations, hydrodynamic equations and their computational implementation are presented in Coco et al. (2007a). A three-dimensional model domain with periodic horizontal boundary conditions is used to represent a seabed composed of two grain sizes ($d_{\mathrm{coarse}} = 0.0005$ m and $d_{\mathrm{fine}} = 0.0002$ m; fall velocity $w_{\mathrm{coarse}} = 0.07\,\mathrm{m\,s^{-1}}$ and $w_{\mathrm{fine}} = 0.02\,\mathrm{m\,s^{-1}}$). An initially flat bed (with slight bathymetric perturbation below 0.01 m) has a bulk composition of 70 % fine sediment and 30 % coarse sediment with individual cells that deviate from this ratio no more than 10 %. The model domain has a plan view size of 500 m × 500 m, a vertical resolution of 0.05 m and a horizontal resolution of 5 m. Small-scale sorted bedforms are modeled in the interest of computational efficiency (observed sorted bedforms range from the scale modeled to kilometers in plan view). In the experiments presented the initial water depth is 9 m, the wave period is 10 s, wave height is 2 m, the mean current is $0.2\,\mathrm{m\,s^{-1}}$, and the current is unidirectional. Sediment transport, computed independently for each size fraction, occurs only as suspended load and results in the change of bed elevation.

Suspended sediment transport is based on a simplified advection–diffusion framework, neglecting horizontal diffusion and assuming steady-state suspended sediment concentration profiles (Murray and Thieler, 2004; Coco et al., 2007a). The flux of suspended sediment ($\boldsymbol{q}_{\mathrm{susp,s}}$), evaluated separately for each size fraction s, is the vertically integrated product of the current velocity profile ($\boldsymbol{V}(z)$) and the suspended sediment concentration profile ($C_{\mathrm{s}}(z)$, where z is the vertical coordinate) combined with a "morphodynamic diffusion" term to incorporate the role of bed slope ($\boldsymbol{\nabla} z$) on sediment transport:

$$\boldsymbol{q}_{\mathrm{susp,s}} = \int C_{\mathrm{s}} \boldsymbol{V} \mathrm{d}z - \gamma_{\mathrm{s}} \frac{1}{5 w_{\mathrm{s}}} U_{\mathrm{w}}^{5} \boldsymbol{\nabla} z, \tag{4}$$

$$\gamma_\mathrm{s} = \gamma_\mathrm{c}\frac{16E\rho}{3\pi w_\mathrm{s}}C_\mathrm{d}, \tag{5}$$

where U_w is the maximum wave orbital speed at the bed ($\mathrm{m\,s^{-1}}$; evaluated with linear wave theory), γ_c is the morphodynamic diffusion coefficient, ρ is the density of water, C_d is the drag coefficient, and E is an efficiency factor (set to 0.035). The integration of suspended sediment flux begins at the height where reference concentration is defined. The second term in Eq. (4) represents a "morphodynamic diffusion" term derived from energetics arguments (Bowen, 1980; Bailard, 1981). The calibration parameter in this framework is γ_c and is adjusted to maintain an order of magnitude difference between the two terms on the right-hand side of Eq. (4), similar to the methodology of Calvete et al. (2001). For all experiments in this contribution, $\gamma_\mathrm{c} = 0.07$. The role of this parameter is addressed further in the discussion section.

Previous work by Coco et al. (2007a) demonstrates negligible sensitivity to different vertical current profile parameterizations (i.e., descriptions that include current–wave interactions). In these experiments we use a logarithmic vertical current profile:

$$V(z) = \frac{1}{\kappa}U^*\log\frac{z}{z_0}. \tag{6}$$

where U^* is the shear velocity and κ is the von Kármán constant. The current profile begins at the roughness height z_0, which is related to wave-generated ripples (van Rijn, 1993):

$$z_0 = \frac{1}{30}\left(2d_{50} + 28\eta\vartheta\right), \tag{7}$$

where η is ripple height and ϑ is ripple steepness.

The wave-period-averaged vertical suspended sediment profile above wave-generated ripples (C_s) is calculated based on Nielsen (1992):

$$C_\mathrm{s}(z) = C_{0,s}e^{-\frac{w_\mathrm{S}z}{\varepsilon_\mathrm{s}}} \tag{8}$$

where $C_{0,s}$ is the near-bed reference concentration for grain size s and ε_s is the vertical sediment diffusivity. Coco et al. (2007a) relied on the formulation developed by Nielsen (1986) to determine the near-bed reference concentration. We use the new GP-derived formulation developed in the previous section. To make the GP-derived C_0 predictor compatible with this model formulation, we assume $U_\mathrm{sig} = U_\mathrm{w}$ and $d_{50} = d_\mathrm{s}$, and therefore Eq. (3) becomes

$$C_0 = \left(\frac{0.328U_\mathrm{w}}{0.0688 + (1000d_\mathrm{s})}\right)^2. \tag{9}$$

The reference concentration is applied at the height of the ripple crest, as in Coco et al. (2007a). In contrast to the work of Coco et al. (2007a) in this work we evaluate the sediment diffusion coefficient based on the work of Nielsen (1992):

$$\varepsilon_\mathrm{s} = \Omega k_s U_\mathrm{w}, \tag{10}$$

$$k_s = 25\eta\vartheta, \tag{11}$$

where k_s is the equivalent roughness and Ω is a scaling coefficient. Thorne et al. (2009) demonstrated that this parameterization underpredicts vertical sediment diffusivity by a factor of ~ 2 when using the original value of $\Omega = 0.016$ suggested by Nielsen (1992). We therefore set $\Omega = 0.032$. Ripple prediction is performed using a new equilibrium scheme developed using GP by Goldstein et al. (2013):

$$\eta = \frac{0.313d_0\left(1000d_{50}\right)}{1.12 + 2.18\left(1000d_{50}\right)}, \tag{12}$$

$$\vartheta = \frac{3.42}{22 + \left(\frac{d_0}{1.12(1000d_{50}) + 2.18(1000d_{50})^2}\right)^2}. \tag{13}$$

We evaluate the mean grain size at each model cell i ($d_{50,i}$) at each time step as

$$d_{50,i} = \left(1 - B_{\mathrm{coarse},i}\right)d_\mathrm{fine} + B_{\mathrm{coarse},i}d_\mathrm{coarse}, \tag{14}$$

where $B_{\mathrm{coarse},i}$ is the percentage of coarse sediment in the active layer at location i, and d_fine and d_coarse are the diameter of the fine and coarse fraction, respectively. An active layer vertically restricts sediment–flow interactions. All experiments presented here have a constant active layer thickness of 0.15 m. Sensitivity analyses performed by Coco et al. (2007a) demonstrate that the nature of the sorting feedback is not changed by modification of the active layer thickness.

5 Hybrid sorted bedform model results

The initially flat, well-mixed conditions can be seen in Fig. 7. This configuration is unstable, and sorted bedforms emerge within 50 model days to form the rhythmic segregated pattern shown in Fig. 7. This self-organization is a consequence of the sorting feedback. Compared to previous modeling, bedforms develop more slowly in the hybrid model. The flux of suspended sediment is smaller for the hybrid model because of the change in reference concentration predictor. Bedforms show an abundance of pattern defects (bifurcations, terminations, and "eyes"), and after initial development the pattern continues to develop through time as a result of bedform interactions: a process of coarsening and pattern maturation occurs as defects move through the system and coarse domains merge to form combined features. This leads to fewer pattern elements (coarse domains) seen through time in Fig. 7. Under unidirectional forcing the sorted bedforms migrate slowly in the direction of the current and profile views show that coarse sediment domains are located along the updrift flank. Fine material is advected downdrift and deposited on the lee side of the coarse domains. Coarse sediment is also transported downdrift, but its mobility is limited on upslope surfaces and in fine domains (where wave-generated bedforms

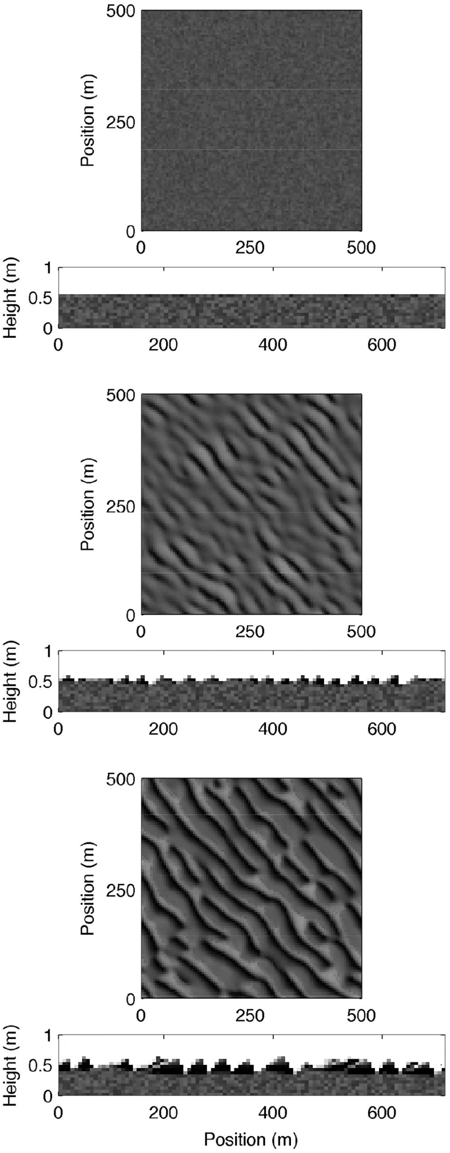

Figure 7. Plan view and profile view of sorted bedform model output (note the vertical exaggeration of profile view). Black and white pixels indicate fine ($d_{\mathrm{fine}} = 0.0002\,\mathrm{m}$) and coarse ($d_{\mathrm{coarse}} = 0.0005\,\mathrm{m}$) sediment, respectively. Current direction is from lower left to upper right and the profile is taken along this axis. The well-mixed and flat initial condition is shown in the top panels. Sorted bedforms appear within 50 days (middle panels) and are well developed by model day 100 (bottom panels). These are mode 2 bedforms; note that coarse domains appear on the updrift flank of the bedforms and wavelength and height are relatively small

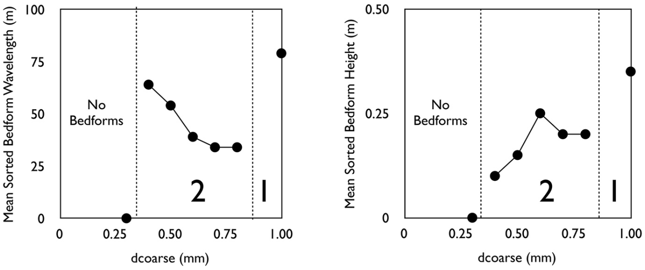

Figure 8. Variations in sorted bedform characteristics (wavelength and height) after 100 days when coarse grain size is held constant. No bedforms appear when the coarse material is too fine. Mode 1 bedforms (long wavelength, larger relief, coarse domains in trough) appear when coarse grain size is large and relatively immobile. Mode 2 bedforms (short wavelength, low relief, coarse domains on updrift flank) appear when coarse grain is between these two limits. No clear pattern was observed after 100 days when $d_{\mathrm{coarse}} = 0.9\,\mathrm{mm}$.

are smaller), and therefore it tends to occupy the updrift flank of the bedform only.

Previous work by Coco et al. (2007a) showed the effect of variations in the size of the fine fraction while the coarse fraction size was held constant. In these experiments we evaluate the reverse: fine fraction diameter is held constant ($d_{\mathrm{fine}} = 0.0002\,\mathrm{m}$; $w_{\mathrm{fine}} = 0.02\,\mathrm{m\,s^{-1}}$), while the coarse fraction diameter is varied between 0.0003 and 0.001 m ($w_{\mathrm{coarse}} = 0.04$–$0.12\,\mathrm{m\,s^{-1}}$). This range of sizes for the coarse fraction is similar to the values found in sorted bedform fields worldwide (Coco et al., 2007b).

Results from this analysis can be seen in Fig. 8 (sorted bedform wavelength and height are evaluated after 100 model days). Similar to Coco et al. (2007a) sorted bedforms do not appear when the grain size contrast between size fractions is too small ($d_{\mathrm{fine}}/d_{\mathrm{coarse}} < 0.5$). When coarse grains range from 0.004 to 0.008 m in diameter, larger coarse sediment tends to cause sorted bedforms to appear faster, decrease in wavelength, and increase in height. Within this range of grain sizes the coarse domain is located along the updrift flank and bedforms migrate in the current direction.

When coarse sediment diameter is larger than 0.008 m, bedforms are strikingly different: bedforms develop faster, wavelengths and height increase significantly, coarse sediment is only present in the trough of the bedform (not along the updrift flank), and bedforms migrate upstream (Fig. 9). This behavior is autogenic in the hybrid sorted bedform model. This pattern configuration is not observed under steady wave climates in the Coco et al. (2007a) model and only appears as the result of specific changes in forcing (Coco et al., 2007b). Bedforms migrate rapidly upcurrent as a result of the decreased mobility of coarse sediment: coarse material is mobile but is not transported significantly up the flank of the bedform and instead remains predominantly in the trough. This is a result of low coarse sediment mobility relative to the downslope transport term in Eq. (4). As fine sediment is advected past the coarse domain in the bedform trough, it can be deposited on the updrift side of the bedform (there is no coarse sediment to prevent its deposition). Along the downdrift side of the bedform the downstream increases in downslope gradient (convex-upward curvature) tends to cause the erosion of bed material and its suspension. This

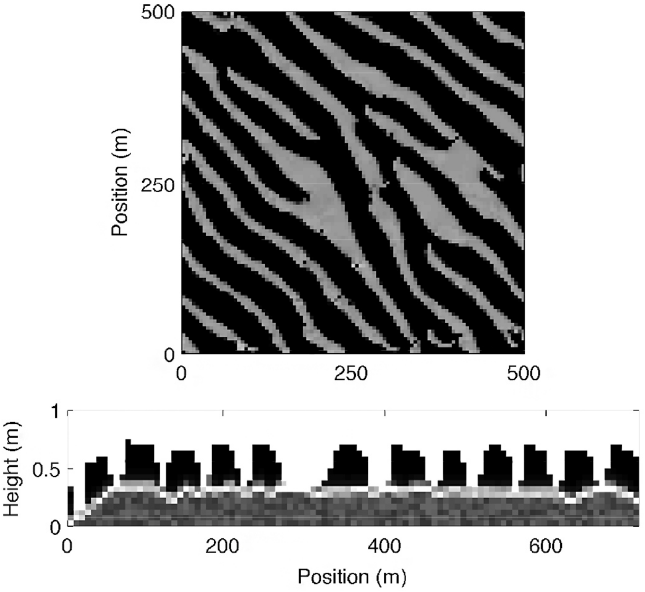

Figure 9. Plan view and profile view of mode 1 sorted bedforms after 50 days. Conditions are identical to Fig. 7 except d_{coarse} = 0.001 m. From identical initial conditions sorted bedforms appear much faster and are prominent features by 50 model days. Note that coarse domains appear solely in the bathymetric trough of the bedforms and wavelength and height are relatively large.

suspended material is advected over the coarse domain (the bedform trough) and subsequently deposited on the updrift side of the following (downdrift) bedform.

In profile view a contiguous layer of coarse sediment exists directly below the sorted bedform field (Fig. 9). This coarse layer occurs at the interface between the well-mixed sediment below (the undisturbed model initial conditions) and the reworked sediment above, a consequence of limited coarse sediment mobility and bedform migration (Goldstein et al., 2011). As bedforms migrate, the position of the sorted bedform trough changes. Fine sediment under the bedform trough, once too deep to experience fluid–sediment interactions, is excavated and suspended. Winnowing of fine sediment and coarsening locally in the bedform trough, repeated as the bedforms migrate, results in the development of a horizontal layer of buried coarse sediment, a “sorting lag”.

In all results presented here, bedforms migrate and bedform wavelength continues to grow through the model run and wavelength does not saturate. This perpetual coarsening of wavelength under conditions of unidirectional currents is identical to the behavior of the Coco et al. (2007b) and Murray and Thieler (2004) model under unidirectional current forcing. (In the previous results, wavelength coarsening also occurs under the more realistic conditions of an asymmetrically reversing current, although coarsening is more gradual than under a unidirectional current.)

6 Discussion

6.1 GP-derived C_0 predictor

The newly developed C_0 predictor has a nonlinear dependence on d_{50} and U_{sig}, similar to other previous empirical predictors (Nielsen, 1986; Lee et al., 2004). This dependence is not imposed, but instead a result of the data sets used in the GP algorithm.

The GP reference concentration predictor relies on U_{sig}, while the sorted bedform model uses U_{w}. In the hybrid model we assume $U_{\mathrm{sig}} = U_{\mathrm{w}}$, where U_{w} is calculated from linear wave theory. We direct the reader to other methods available to estimate U_{sig} from surface wave parameters (e.g., Wiberg and Sherwood, 2008). We force the sorted bedform model with a constant monochromatic wave field (height and period) to eliminate the chance that changes in wave characteristics influence the simulated seabed evolution. Therefore the assumption of $U_{\mathrm{sig}} = U_{\mathrm{w}}$ does not impact model results shown here.

Ripple geometry was not used as an independent variable in the construction of the C_0 predictor. Dolphin and Vincent (2009) recently suggested that ripple geometry may not aid in the prediction of C_0, contrary to Nielsen (1986) and Green and Black (1999). Though we do not have data to either support or refute this claim, we can offer our results as an example of a well-performing prediction of reference concentration without the explicit inclusion of ripple geometry. However, the nonlinear nature of the reference concentration prediction and the constants embedded within Eq. (3) suggest that ripple configuration may be encoded within the predictor, either as a cause of the nonlinearity or a determinant of the constants.

The C_0 predictor does not explicitly account for near-bed currents that may be important mechanisms for enhancing suspension in sorted bedform fields (e.g., Gutierrez et al., 2005). The C_0 predictor developed in this study is an equilibrium predictor; therefore the role of time variance of C_0 is not addressed (e.g., Vincent and Hanes, 2002). However, the data were collected in burst mode, a technique that involves time averaging. Burst measurements may reduce the effect of some time-dependent processes (e.g., advected clouds of sediment, wave groups, etc.). The GP predictor is constructed solely with regard to the measurement data and is not based on “first principles”. Using the independent testing data, the new GP predictor has a lower NRMSE and identical correlation coefficient than the Nielsen (1986) and Lee et al. (2004) predictors; however the GP predictor does not perform well at low concentrations (Fig. 6). The poor performance may be the result of nonlinearities in sediment transport that are not captured by the prediction scheme, noise in the experimental signal at low concentrations, or other as yet unknown reasons. Notably, more energetic conditions are required to move sediment using the GP predictor than compared to the Nielsen (1986) prediction scheme previously used in the sorted bedform model. This result is similar to previous work that suggests the Nielsen (1986) predictor may overestimate reference concentration (Bolaños et al., 2012; Thorne et al., 2002).

6.2 Hybrid sorted bedform model

The hybrid version of the sorted bedform model is able to reproduce the sorting feedback using new parameterizations built from data. The sorting feedback hypothesized by Murray and Thieler (2004) is robust to changes in the mathematical description of the processes in sediment transport and hydrodynamics on the continental shelf, and hybrid model results are comparable to previous modeling efforts (Murray and Thieler, 2004; Murray et al., 2005; Coco et al., 2007a). The behavior of the hybrid model and the Coco et al. (2007a) model under identical hydrodynamic forcing is different because there are quantitative differences between the mathematical description of sediment transport processes. For instance, using the baseline conditions of the Coco et al. (2007a) model the hybrid model produces no sorted bedforms. This is a direct result of changing the C_0 predictor from the Nielsen (1986) formula (which overpredicts sediment transport; Fig. 6) to the new GP-derived C_0 predictor. Changes to the sediment transport formulas prohibit us from directly comparing the three models under identical forcing conditions. Instead we offer this hybrid model as a refined version of the Coco et al. (2007a) model. The hybrid model has additional advantages beyond being more tightly coupled to observational data, most notably in favorable comparison to previous observational work.

Results shown in this contribution use two new prediction schemes based on GP (i.e., ripple morphology and reference concentration). We believe the new ripple prediction scheme of Goldstein et al. (2013) is an improvement over the previous method used in the Coco et al. (2007a) model; however ripples in this model only significantly impact the vertical sediment diffusivity (ε_s) and the roughness height (z_0). The reference concentration, since it sets the magnitude of suspended sediment, is more strongly related to the new behaviors in the model, and as a result we focus our analysis on the reference concentration.

Observational work has previously detected several distinct varieties of sorted bedforms – those with coarse sediment in the trough and those where coarse sediment appears either in the trough and bedforms where coarse sediment is located on the flank (both the updrift and/or downdrift; e.g., Goff et al., 2005; Ferrini and Flood, 2005). Van Oyen et al. (2010, 2011) found that these two pattern configurations appear in linear stability analysis as a result of two separate feedback mechanisms. Mode 1 bedforms (flow–topography feedback), where coarse domains are located in the bedform trough, have a faster growth rate when waves and currents are weaker and result in bedforms with longer wavelength, larger amplitude, and faster migration rates. Mode 2 bedforms (sorting or "roughness" feedback), where coarse grains appear along the updrift and downdrift flank of the bedform, have a faster growth rate when waves and currents are stronger and result in bedforms with smaller wavelengths, smaller heights, and slower migration rates. Yet results from linear stability analysis are applicable only at the scale of an infinitesimal perturbation.

Results from the finite-amplitude hybrid model also show that coarse domains can occur either on the updrift flank of the sorted bedform or collocated with the bedform trough, matching some aspects of previous observation work. However instead of relying on two separate feedback mechanisms, the hybrid model is able to reproduce these two pattern configurations solely via the sorting mechanism. The presence of two distinct pattern modes occurs while current and wave conditions remain unchanged but coarse grain size is varied. When coarse grains are smaller (essentially identical to increasing wave conditions in terms of increasing coarse sediment mobility) bedforms conform to the description of the mode 2 features of Van Oyen et al. (2010, 2011) with smaller features, slower migration rates, and coarse sediment along the updrift flank of bedforms. When coarse grains are larger (essentially identical to decreasing wave conditions in terms of decreasing coarse sediment mobility) bedforms show characteristics of the mode 1 features of Van Oyen et al. (2010, 2011) with larger bedforms, faster migration rates, and coarse sediment in the bedform trough. We again note this behavior occurs solely from a sorting feedback. Bedform wavelength continues to grow in all model results shown here as a result of unidirectional current. However, results in this contribution show that, for any given instant in model time, modeled sorted bedform patterns display relatively homogenous wavelength and height (similar to Coco et al. (2007a) and Murray and Thieler (2004)). Observational work shows sorted bedform fields have a well-defined pattern scale (i.e., a similar height and wavelength throughout the entire bedform field; see the compilation of observed bedform features in Coco et al. (2007b) for more details). It remains unknown whether the well-defined pattern scale of observed sorted bedforms reflects a saturated (steady state) wavelength or the uniformity of bedform wavelength and height at a given moment of pattern evolution.

Several features of mode 1 bedforms in the hybrid model warrant additional attention. Linear stability analysis (Van Oyen et al., 2010, 2011) suggests infinitesimal mode 1 bedforms should migrate in the current direction. The large-scale mode 1 bedforms formed in the finite-amplitude hybrid model show upcurrent migration, which has not previously been observed in field examples of sorted bedforms. Furthermore, mode 1 bedforms develop in the linear stability analysis as a result of a flow–bathymetry feedback (Van Oyen et al., 2010, 2011). The finite-amplitude hybrid model presented here does not parameterize hydrodynamics at small enough scales to permit the development of bedforms as a result of a flow–bathymetry feedback. In contrast to the linear stability analysis, mode 1 bedforms in the hybrid model develop as result of the sorting feedback operating at finite amplitude. Future work with more detailed hydrodynamic parameterizations could shed light on the interplay between flow–bathymetry interactions and the sorting feedback in the mode 1 regime at finite amplitudes. However, these results do suggest that the finite-amplitude hybrid model may be able to capture the dynamics observed in the field. The presence of two distinct pattern modes in the hybrid model is a direct result of incorporating new data-driven parameterizations of the sediment transport process. In this contribution we explore only one specific mechanism that results in mode 1 sorted bedforms, increasing the diameter of the coarse grain size fraction. There are likely other mechanism by which mode 1 bedforms may develop instead of mode 2 bedforms, notably by increasing water depth, decreasing wave forcing, or decreasing current velocity.

There are additional pattern-scale consequences to adjusting the sediment transport formulations. The new C_0 predictor requires energetic conditions to move coarse sediment. This matches the observations and interpretations of Green et al. (2004), Trembanis et al. (2004), and Trembanis and Hume (2011), who suggest that energetic conditions are the only time when the coarse sediment of sorted bedforms is mobile. However lower coarse sediment mobility results in the creation of more pattern defects, a common feature of field examples of sorted bedforms (e.g., Fig. 1). Furthermore, after the work of Werner and Kocurek (1997, 1999), defects have been recognized as a fundamental variable in pattern-scale dynamics of bedforms (Huntley et al., 2008; Maier and Hay, 2009; Goldstein et al., 2011; Skarke and Trembanis, 2011). The presence of additional defects in the hybrid model may exert fundamental controls on pattern evolution.

The hybrid model is able to reproduce sorting feedback and two pattern modes when successfully calibrated. Calibration is accomplished by adjusting the variable γ_c in the morphodynamic diffusion term, Eqs. (4) and (5). The results shown in this contribution have $\gamma_c = 0.07$. The sorting feedback and the development of two sorted bedform pattern modes occur in the range of $\gamma_c = 0.05$–0.08. This range contrasts with the work of Coco et al. (2007a, b), where the γ_c term could be adjusted at least one order of magnitude. This more limited calibration is the result of using multiple nonlinear elements in the construction of the model. Specifically the morphodynamic diffusion term (that γ_c modifies) is highly nonlinear (i.e., $\propto U_\mathrm{w}^5$) and is built from energy-based theory (Bowen, 1980; Bailard, 1981). Coco et al. (2007a) relied on a parameterization of C_0 that scaled with U_w^6, effectively scaling the two terms of Eq. (4) in a similar manner. In contrast our new C_0 predictor scales with U_w^2, and therefore does not scale in a similar manner to the morphodynamic term (U_w^5). We suggest that this mismatch, coupled with the strong forcing condition that is required to move sediment in the model (i.e., large U_w), has lead to a smaller permissible parameter space where the morphodynamic term and the new GP derived predictor are interoperable. We define the permissible parameter space by the scaling argument made previously by Calvete et al. (2001): γ_c should be set to a value that maintains the ratio between the two terms on the right side of Eq. (4) to ∼ 1 order of magnitude. If γ_c is set too high, the slope-dependent term is too strong and no bathymetric perturbations develop. If γ_c is set too low, nonphysically steep bathymetric perturbations develop. These results highlight the need to test the Bailard (1981) term in a range of conditions to see whether this description (or others) is valid. Though this morphodynamic diffusion term is often used in morphodynamic models, we could find no instance where this term has been tested in a wide range of conditions.

Finally, the promising results of data-driven parameterizations as components in the sorted bedform model suggests that this approach could be extended to other morphodynamic models and other parameterizations. A specific example from this work is the parameterization of vertical sediment diffusivity (or, more generally, the shape function that described the vertical suspended sediment concentration profile). Recent work has begun to investigate the fast scale dynamics of vertical sediment diffusion over ripples (e.g., Davies and Thorne, 2005; van der Werf et al., 2007; O’Hara Murray et al., 2011) and how best to parameterize this process in large-scale coastal models (Amoudry and Souza, 2011; Amoudry et al., 2013). Traditional equilibrium parameterizations have also been evaluated with newly collected data (e.g., Thorne et al., 2002, 2009; Bolaños et al., 2012). More data, collected in a range of conditions, would enable a data-driven approach to the parameterization of the vertical suspended sediment profile shape.

7 Conclusion

A new predictor for near-bed reference concentration developed using genetic programming performs as well or better than previous empirical parameterizations. However the GP predictor shows poor performance at low concentrations. This predictor is incorporated, along with previously developed predictors for ripple morphology (developed by GP), into a new “hybrid” model of sorted bedforms. This modeling strategy is a viable option when large data sets can be used to construct data-driven subcomponents of a morphodynamic model. The sorting feedback is relatively invariant to changes in hydrodynamic and sediment transport parameterizations. However, the new hybrid model is able to generate novel autogenic behavior in the sorted bedform model: sorted bedform morphology changes when the size of the coarse fraction is modified. This model behavior more closely resembles field observations showing sorted bedform coarse domains that occur in multiple positions along the bedform (however downdrift coarse domains still do not appear in this model)

Acknowledgements. We thank Paula Camus for sharing her clustering algorithm and to Terry Hume (NIWA, NZ) for providing the image of Tairua Beach, NZ, bedforms seen in Fig. 1. E. B. Goldstein thanks "IH Cantabria" for funding during his stay, where this work was started. G. Coco acknowledges funding from the Cantabria Campus Internacional (Augusto Gonzalez Linares program). We thank an anonymous reviewer and Tomas Van Oyen for stimulating comments that improved the manuscript.

Edited by: F. Metivier

References

Amoudry, L. O. and Souza, A. J.: Deterministic coastal morphological and sediment transport modeling: A review and discussion, Rev. Geophys., 49, RG2002, doi:10.1029/2010RG000341, 2011.

Amoudry, L. O., Bell, P. S., Thorne, P. D., and Souza, A. J.: Toward representing wave-induced sediment suspension over sand ripples in RANS models, J. Geophys. Res.-Oceans, 118, 1–15, 2013.

Bailard, J. A.: An energetics total load sediment transport model for a plane sloping beach, J. Geophys. Res., 86, 10938–10954, 1981.

Baptist, M. J., Babovic, V., Uthurburu, J. R., Keijzer, M., Uittenbogaard, R. E., Mynett, A., and Verwey, A.: On inducing equations for vegetation resistance, J. Hydraul. Res., 45, 435–450, 2007.

Bolaños, R., Thorne, P. D., and Wolf, J.: Comparison of measurements and models of bed stress, bedforms and suspended sediments under combined currents and waves, Coast. Eng., 62, 19–30, 2012.

Bowden, G. J., Maier, H. R., and Dandy, G. C.: Optimal division of data for neural network models in water resources applications, Water Resour. Res., 38(2), 2-1–2-11, 2002.

Bowen, A. J.: Simple models of nearshore sedimentation: Beach profiles and long-shore bars, in: The Coastline of Canada, edited by: McCann, S. B., 111 pp., Geol. Surv. of Can., Ottawa, 1980.

Cacchione, D. A., Thorne, P. D., Agrawal, Y., and Nidzieko, N. J.: Time-averaged near-bed suspended sediment concentrations under waves and currents: comparison of measured and model estimates, Cont. Shelf Res., 28, 470–484, 2008.

Calvete, D., Falqués, A., de Swart, H. E., and Walgreen, M.: Modelling the formation of shoreface-connected sand ridges on storm-dominated inner shelves, J. Fluid Mech., 441, 169–193, 2001.

Camus, P., Mendez, F. J., Medina, R., and Cofiño, A. S.: Analysis of clustering and selection algorithms for the study of multivariate wave climate, Coast. Eng., 58, 453–462, 2011.

Coco, G., Murray, A. B., and Green, M. O.: Sorted bedforms as self-organized patterns: 1. Model development, J. Geophys. Res., 112, F03015, doi:10.1029/2006JF000665, 2007a.

Coco, G., Murray, A. B., Green, M. O., Thieler, E. R., and Hume, T. M.: Sorted bedforms as self-organized patterns: 2. Complex forcing scenarios, J. Geophys. Res., 112, F03016, doi:10.1029/2006JF000666, 2007b.

Corzo, G. A., Solomatine, D. P., Hidayat, de Wit, M., Werner, M., Uhlenbrook, S., and Price, R. K.: Combining semi-distributed process-based and data-driven models in flow simulation: a case study of the Meuse river basin, Hydrology and Earth System Sciences, 13, 9, 1619–1634, 2009.

Cummings, D. I., Dumas, S., and Dalrymple, R. W.: Fine-grained versus coarse-grained wave ripples generated experimentally under large-scale oscillatory flow, J. Sediment. Res., 79, 83–93, 2009.

Davies, A. G. and Thorne, P. D.: Modeling and measurement of sediment transport by waves in the vortex ripple regime, J. Geophys. Res., 110, C05017, doi:10.1029/2004JC002468, 2005.

Dolphin, T. and Vincent, C. E.: The influence of bed forms on reference concentration and suspension under waves and currents, Cont. Shelf Res., 29, 424–432, 2009.

Ferrini, V. L. and Flood, R. D.: A comparison of rippled scour depressions identified with multibeam sonar: Evidence of sediment transport in inner shelf environments, Cont. Shelf Res., 25, 1979–1995, 2005.

Gagné, C., Schoenauer, M., Parizeau, M., and Tomassini, M.: Genetic programming, validation sets, and parsimony pressure, in: Genetic Programming, 9th European Conference, EuroGP2006, Lecture Notes in Computer Science, LNCS 3905, edited by: Collet, P., Tomassini, M., Ebner, M., Gustafson, S., and Ekárt, A., Springer, Berlin, Heidelberg, New York, 2006, 109–120, 2006.

Goff, J. A., Mayer, L. A., Traykovski, P., Buynevich, I., Wilkens, R., Raymond, R., Glang, G., Evans, R. L., Olson, H., and Jenkins, C.: Detailed investigation of sorted bedforms, or "rippled scour depressions", within the Martha's Vineyard Coastal Observatory, Massachusetts, Cont. Shelf Res., 25, 461–484, 2005.

Goldstein, E. B., Murray, A. B., and Coco, G.: Sorted bedform pattern evolution: Persistence, destruction and self-organized intermittency, Geophys. Res. Lett., 38, L24402, doi:10.1029/2011GL049732, 2011.

Goldstein, E. B., Coco, G., and Murray, A. B.: Prediction of wave ripple characteristics using genetic programming, Cont. Shelf Res., 71, 1–15, 2013.

Gonçalves, I. and Silva, S.: Balancing Learning and Overfitting in Genetic Programming with Interleaved Sampling of Training Data, in: EuroGP 2013. LNCS, edited by: Krawiec, K., Moraglio, A., Hu, T., Etaner-Uyar, A., and Hu, B., Springer, Heidelberg, 7831, 73–84, 2013.

Gonçalves, I., Silva, S., Melo, J., and Carreiras, J.: Random Sampling Technique for Overfitting Control in Genetic Programming, in: EuroGP 2012. LNCS, edited by: Moraglio, A., Silva, S., Krawiec, K., Machado, P., and Cotta, C., Springer, Heidelberg, 7244, 218–229, 2012.

Green, M.: Introducing ALICE, Water and Atmosphere, 4, 8–10, 1996.

Green, M. O.: Test of sediment initial-motion theories using irregular-wave field data, Sedimentology, 46, 427–441, 1999.

Green, M. O. and Black, K. P.: Suspended-sediment reference concentration under waves: field observations and critical analysis of two predictive models, Coast. Eng., 38, 115–141, 1999.

Green, M. O. and MacDonald, I. T.: Processes driving estuary infilling by marine sands on an embayed coast, Mar. Geol., 178, 11–37, 2001.

Green, M. O., Vincent, C. E., and Trembanis, A. C.: Suspension of coarse and fine sand on a wave-dominated shoreface, with implications for the development of rippled scour depressions, Cont. Shelf Res., 24, 317–335, 2004.

Gutierrez, B. T., Voulgaris, G., and Thieler, E. R.: Exploring the persistence of sorted bedforms on the inner-shelf of Wrightsville Beach, North Carolina, Cont. Shelf Res., 25, 65–90, 2005.

Huntley, D. A., Coco, G., Bryan, K. R., and Murray, A. B.: Influence of "defects" on sorted bedform dynamics, Geophys. Res. Lett., 35, L02601, doi:10.1029/2007GL030512, 2008.

Karunarathna, H. and Reeve, D. E.: A hybrid approach to model shoreline change at multiple timescales, Cont. Shelf Res., 66, 29–35, 2013.

Kitsikoudis, V., Sidiropoulos, E., and Hrissanthou, V.: Derivation of Sediment Transport Models for Sand Bed Rivers from Data-Driven Techniques, in: Sediment Transport Processes and Their Modelling Applications, edited by: Manning, A., ISBN: 978-953-51-1039-2, InTech, doi:10.5772/53432, 2013.

Knaapen, M. A. F. and Hulscher, S. J. M. H. : Regeneration of sand waves after dredging, Coast. Eng., 46, 277–289, 2002.

Knaapen, M. A. F. and Hulscher, S. J. M. H.: Use of a genetic algorithm to improve predictions of alternate bar dynamics, Water Resour. Res., 39, 1231, doi:10.1029/2002WR001793, 2003.

Koza, J. R.: Genetic Programming, On the Programming of Computers by Means of Natural Selection, MIT Press, Cambridge, MA, USA, 1992.

Krasnopolsky, V. M. and Fox-Rabinovitz, M. S.: A new synergetic paradigm in environmental numerical modeling: Hybrid models combining deterministic and machine learning components, Ecol. Model., 191, 5–18, 2006.

Kushchu, I.: An evaluation of evolutionary generalisation in genetic programming, Artif. Intell. Rev., 18, 3–14, 2002.

Lee, G., Dade, W. B., Friedrichs, C. T., and Vincent, C. E.: Examination of reference concentration under waves and currents on the inner shelf, J. Geophys. Res., 109, C02021, doi:10.1029/2002JC001707, 2004.

Maier, I. and Hay, A. E.: Occurrence and orientation of anorbital ripples in near-shore sands, J. Geophys. Res., 114, F04022, doi:10.1029/2008JF001126, 2009.

May, R. J., Maier, H. R., and Dandy, G. C.: Data splitting for artificial neural networks using SOM-based stratified sampling, Neural Networks, 23, 283–294, 2010.

Murray, A. B. and Thieler, E. R.: A new hypothesis and exploratory model for the formation of large-scale inner-shelf sediment sorting and "rippled scour depressions", Cont. Shelf Res., 24, 295–315, 2004.

Murray, A. B., Coco, G., Green, M. O., Hume, T., and Thieler, E. R.: Different approaches to modeling inner shelf sorted bedforms, in: Proceedings of the Conference "River, Coastal and Estuarine Morphodynamics", edited by: Parker, G. and Garcia, M., Taylor and Francis, London, 1009–1015, 2005.

Nielsen, P.: Suspended sediment concentrations under waves, Coast. Eng., 10, 23–31, 1986.

Nielsen, P.: Coastal Bottom Boundary Layers and Sediment Transport, World Sci., Singapore, 1992.

Oates, T. and Jensen, D.: The effects of training set size on decision tree complexity, Proceedings of the Fourteenth International Conference on Machine Learning, Madison, WI, Morgan Kaufmann, 254–262, 1997.

Oates, T. and Jensen, D.: Large datasets lead to overly complex models: An explanation and a solution, Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, New York, NY, AAAI Press, 294–298, 1998.

Oehler, F., Coco, G., Green, M. O., and Bryan, K. R.: A data driven approach to predict suspended-sediment reference concentration under non-breaking waves, Cont. Shelf Res., 46, 96–106, 2012.

O'Hara Murray, R. B., Thorne, P. D., and Hodgson, D. M.: Intrawave observations of sediment entrainment processes above sand ripples under irregular waves, J. Geophys. Res., 116, C01001, doi:10.1029/2010JC006216, 2011

O'Neill, M., Vanneschi, L., Gustafson, S., and Banzhaf, W.: Open issues in genetic programming, Genet. Program. Evol. M., 11, 339–363, 2010.

Panait, L. and Luke, S.: Methods for Evolving Robust Programs, in: Genetic and Evolutionary Computation-GECCO 2003, Vol. 2724, no. 66, Berlin, Heidelberg, Springer Berlin Heidelberg, 1740–175, 2003.

Pape, L., Ruessink, B. G., Wiering, M. A., and Turner, I. L.: Recurrent neural network modeling of nearshore sandbar behavior, Neural Networks, 20, 509–518, 2007.

Pape, L., Kuriyama, Y., and Ruessink, B. G.: Models and scales for cross-shore sandbar migration, J. Geophys. Res., 115, F03043, doi:10.1029/2009JF001644, 2010.

Poli, R., Langdon, W. B., and McPhee, N. F.: A field guide to genetic programming, Lulu Enterprises Uk Limited, 2008

Ruessink, B. G.: Calibration of nearshore process models: Application of a hybrid genetic algorithm, J. Hydroinformatics, 7, 135–149, 2005.

Schmidt, M. and Lipson, H.: Distilling free-form natural laws from experimental data, Science, 324, 81–85, 2009.

Schmidt, M. and Lipson, H.: Eureqa (Version 0.98 beta) [Software], available at: http://www.eureqa.com/, 2013.

Skarke, A. and Trembanis, A. C.: Parameterization of bedform morphology and defect density with fingerprint analysis techniques, Cont. Shelf Res., 31, 1688–1700, 2011.

Thieler, E. R. Foster, D. S., Himmelstoss, E. A., and Mallinson, D. J.: Geologic framework of the northern North Carolina, USA inner continental shelf and its influence on coastal evolution, Mar. Geo., 348, 113–130, 2014.

Thorne, P. D., William, J. J., and Davies, A. G.: Suspended sediments under waves measured in a large scale flume facility, J. Geophys. Res., 107, 4.1–4.16, 2002.

Thorne, P. D., Davies, A. G., and Bell, P. S.: Observations and analysis of sediment diffusivity profiles over sandy rippled beds under waves, J. Geophys. Res., 114, C02023, doi:10.1029/2008JC004944, 2009.

Trembanis, A. C. and Hume, T. M.: Sorted bedforms on the inner shelf off northeastern New Zealand: spatiotemporal relationships and potential paleo-environmental implications, Geo-Mar Letters, 31, 203–214, 2011.

Trembanis, A. C., Wright, L. D., Friedrichs, C. T., Green, M. O., and Hume, T.: The effects of spatially complex inner shelf roughness on boundary layer turbulence and current and wave friction: Tairua embayment, New Zealand, Cont. Shelf Res., 24, 1549–1571, 2004.

van der Werf, J. J., Doucette, J. S., Donoghue, T., and Ribberink, J. S.: Detailed measurements of velocities and suspended sand concentrations over full?scale ripples in regular oscillatory flow, J. Geophys. Res., 112, F02012, doi:10.1029/2006JF000614, 2007.

van Maanen, B., Coco, G., Bryan, K. R., and Ruessink, B. G.: The use of artificial neural networks to analyze and predict alongshore sediment transport, Nonlin. Processes Geophys., 17, 395–404, 2010,
http://www.nonlin-processes-geophys.net/17/395/2010/.

Van Oyen, T., de Swart, H. E., and Blondeaux, P.: Bottom topography and roughness variations as triggering mechanisms to the formation of sorted bedforms, Geophys. Res. Lett., 37, L18401, doi:10.1029/2010GL043793, 2010.

Van Oyen, T., de Swart, H. E., and Blondeaux, P.: Formation of rhythmic sorted bed forms on the continental shelf: an idealised model, J. Fluid Mech., 684, 475–508, 2011.

van Rijn, L. C.: Principles of Sediment Transport in Rivers, Estuaries and Coastal Seas, Aqua, Amsterdam, 1993.

Vincent, C. E. and Green, M. O.: The control of resuspension over megaripples on the continental shelf, in: Proceedings of Coastal Sediments 99, New York, USA, ASCE, 269–280, 1999.

Vincent, C. E. and Hanes, D. M.: The accumulation and decay of near-bed suspended sand concentration due to waves and wave groups, Cont. Shelf Res., 22, 1987–2000, 2002.

Werner, B. T. and Kocurek, G.: Bedform dynamics: Does the tail wag the dog?, Geology, 25, 771–774, 1997.

Werner, B. T. and Kocurek, G.: Bedform spacing from defect dynamics, Geology, 27, 727–730, 1999.

Wiberg, P. L. and Sherwood, C. R.: Calculating wave-generated bottom orbital velocities from surface-wave parameters, Comput. Geosci., 34, 1243–1262, 2008.

Yoon, H.-D., Cox, D. T., and Kim, M.: Prediction of time-dependent sediment suspension in the surf zone using artificial neural network, Coast. Eng., 71, 78–86, 2013.