Inner Shelf Sorted Bedforms: Long-Term Evolution and a New Hybrid Model

by

Evan Benjamin Goldstein

Department of Earth and Ocean Sciences
Duke University

Date:_______________________
Approved:

____
A. Brad Murray, Supervisor

____
Giovanni Coco

____
Peter K. Haff

____
James B. Heffernan

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of Philosophy in
Earth and Ocean Sciences in the Graduate School
of Duke University

2014

ABSTRACT

Inner Shelf Sorted Bedforms: Long-term evolution and a new 'hybrid' model

by

Evan Benjamin Goldstein

Earth and Ocean Sciences
Duke University

Date:________________________
Approved:

____
A. Brad Murray, Supervisor

____
Giovanni Coco

____
Peter K. Haff

____
James B. Heffernan

An abstract of a dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor of Philosophy in
Earth and Ocean Sciences in the Graduate School
of Duke University

2014

Abstract

Sorted bedforms are spatial extensive (100 m-km) features present on many inner continental shelves with subtle bathymetric relief (cm-m) and localized, abrupt variations in grain size (fine sand to coarse sand/gravel). Sorted bedforms provide nursery habitat for fish, are a control on benthic biodiversity, function as sediment reservoirs, and influence nearshore waves and currents. Research suggests these bedforms are a consequence of a sediment sorting feedback as opposed to the more common flow-bathymetry interaction. This dissertation addresses three topics related to sorted bedforms: 1) Modeling the long-term evolution of bedform patterns, 2) Refinement of morphological and sediment transport relations used in the sorted bedform model with 'machine learning'; 3) Development of a new sorted bedform model using these new 'data-driven' components.

Chapter 1 focuses on modeling the long term evolution of sorted bedforms. A range of sorted bedform model behaviors is possible in the long term, from pattern persistence to spatial-temporal intermittency. Vertical sorting (a result of pattern maturation processes) causes the burial of coarse material until a critical state of seabed coarseness is reached. This critical state causes a local cessation of the sorting feedback, leading to a self-organized spatially intermittent pattern, a hallmark of observed sorted bedforms. Various patterns emerge when numerical experiments include erosion, deposition, and storm events.

Modeling of sorted bedforms relies on the parameterization of processes that lack deterministic descriptions. When large datasets exist, machine learning (optimization tools from computer science) can be used to develop parameterizations directly from data. Using genetic programming (a machine learning technique) and large multisetting datasets I develop smooth, physically meaningful predictors for ripple morphology (wavelength, height, and steepness; Chapter 2) and near bed suspended sediment reference concentration under unbroken waves (Chapter 3). The new predictors perform better than existing empirical formulations.

In Chapter 3, the new components derived from machine learning are integrated into the sorted bedform model to create a ‘hybrid’ model: a novel way to incorporate observational data into a numerical model. Results suggest that the new hybrid model is able to capture dynamics absent from previous models, specifically, the two observed end-member pattern modes of sorted bedforms (i.e., coarse material on updrift bedform flanks or coarse material in bedform troughs). However, caveats exist when data driven components do not have parity with traditional theoretical components of morphodynamic models, and I address the challenges of integrating these disparate pieces and the future of this type of ‘hybrid’ modeling.

Dedication

To Max, Emily, Mom and Dad

Contents

List of Tables

List of Figures

Acknowledgements

I first thank my advisor, A. Brad Murray, for his constant support and guidance. I thank Giovanni Coco for his energy and enthusiasm. I thank Peter Haff for many long conversations. I thank Jim Heffernan and Marco Marani for serving on my committee at the end, and near the end (respectively). I thank Laura Moore (UNC- Chapel Hill) and Dylan McNamara (UNC- Wilmington) for serving on my committee during my preliminary exam. I thank Debbie Gooch, Beatriz Martin, and Debra Colpitts for support. I thank Malcolm Green for sharing his data (and an office). I thank Paula Camus for sharing her MDA routine. I thank my fellow lab mates Patrick Limber, Kenneth Ells, Katherine Murray, and Rebecca Lauzon for camaraderie.

I thank Janice P. Burke and Neil W. Goldstein for unfailing emotional support. I thank Emily M. Janke for her unfailing belief in me, and her long hours of listening. Lastly I thank Maximilian B. Janke for reminding me that, rain or shine, we should go play outside.

1. Sorted Bedform Pattern Evolution: Persistence, Destruction and Self-Organized Intermittency

Goldstein, E. B., A. B. Murray, and G. Coco, 2011, Sorted bedform pattern evolution: Persistence, destruction and self-organized intermittency, Geophysical Research Letters, 38, L24402.

1.1 Introduction

Recent observations have shown that the inner shelf, the dynamic region that links the surf zone and the continental shelf, displays various types of bathymetric features. These features exert a first order control on the role of the inner shelf as a conduit for wave energy and sediment. Here we examine one such feature, spatially extensive (100 m - km scale) sequences of strikingly sorted coarse- and fine-grained sandy domains with slight bathymetric relief (cm - m scale) that appear commonly on the inner continental shelf (Figure 1). Recent research suggests that these "sorted bedforms" arise primarily from a sediment sorting feedback, rather than interactions between flow and topography (Murray and Thieler, 2004, Coco et al., 2007a, Van Oyen et al. 2010). This sorting process is mediated by wave-generated ripples, whose size is a function of bed composition. Areas mantled by coarse sediment exhibit larger ripples than regions covered with fine sediment. Large wave generated ripples lead to strong, large-scale turbulence above coarse domains. Strong turbulence inhibits the settling of suspended fine sediment, leading to preferential fine sediment deposition in locations with less turbulence (i.e. finer sediment). Modeling by Coco et al. (2007a), which builds on previous work by Murray and Thieler (2004), shows that this sorting feedback is robust under a wide range of wave and current forcing conditions.

This previous work addressed sorting dynamics and pattern development on short time scales (days to weeks; Coco et al., 2007a, 2007b). Over longer time scales, observational studies show that sorted bedforms may be temporally intermittent or persistent inner shelf features. Surveys of shallow water environments tend to show ephemeral sorted bedforms while bedforms in deep water are persistent [e.g. Hume et al., 2003; Details in auxiliary material].

Stable sorted bedform patterns in nature differ in form from the highly organized patterns developed in previous modeling endeavors. Observed sorted bedform patterns often exhibit a spatially complicated plan view pattern with ragged, 'wispy' down-drift edges of coarse domains (Murray and Thieler 2004) and many pattern imperfections (Figure 1). These imperfections, termed defects, are the ends of coarse domains or areas where one coarse patch bifurcates into two parallel coarse domains. In this contribution we assimilate these observations into a cohesive picture of long-term sorted bedform dynamics and investigate what role forcing conditions and recurring storm events have in generating spatial intermittence and temporal intermittence, and in determining the degree of organization of the pattern.

To investigate the long-term dynamics of sorted bedforms we use the numerical model developed by Coco et al. (2007a). Long-term pattern evolution is also examined under conditions of net deposition (representing a scenario such as a prograding shoreface associated with a convergence of alongshore sediment flux or a fluvial sediment source), net erosion (representing e.g. scour by strong tidal currents and waves), as well as periodic high wave events (representing storms). We show that the key to a durable sorted bedform field is a sufficiently high concentration of coarse material on the seabed to drive the sorting feedback. Changes in the concentration of coarse material on the seabed results from the self-organized dynamics of the system. The purpose of the study is to explore the long-term morphodynamics under simplified conditions to maximize potential insights in this early stage of investigation rather than reproduce realistic scenarios with maximum accuracy.

1.2 Methods

Here we review the main attributes of the numerical model, which represents the transport of coarse and fine grain sediment as both bedload and suspended load under variable current and wave forcing (see Coco et al. (2007a) for a full model description). Sediment flux is treated in an advection-diffusion framework, neglecting diffusion, with terms for sediment sources (entrainment; a function of wave orbital speed) and sinks (deposition; proportional to suspended sediment concentration) which correspond to changes in bed elevation.

The three dimensional model domain (a plan view size of 500 m x 500 m) has a 5 m horizontal resolution, 0.05 m vertical resolution and periodic horizontal boundary conditions. The initial condition is an approximately flat bed (bathymetric perturbations below 0.01 m) with a bulk composition of 30% coarse sediment (d_{coarse}=0.001 m) and 70% fine sediment (d_{fine}=0.00015 m) with deviations in individual cell coarse percentages below 10%. In these experiments the current reverses direction every 24 hours. Unless otherwise noted mean current velocity is 0.2 m/s (with a daily excursion about the mean current with a standard deviation of 0.06), wave height is 2 m, wave period is 10 s and the initial water depth is 20 m. An active layer limits the vertical depth over which sediment interacts with the flow. For all experiments in this treatment the active layer is a constant thickness of 0.15 m (further model details are in the auxiliary material) Coco et al. (2007b) discussed the role of current and wave variations in plan view pattern development. To examine the influence of higher wave events (storms) on pattern robustness, we include storm scenarios with 4 m waves that occur for 4 days every 100 days (scenarios with increased current magnitudes are discussed in auxiliary material). We are modeling relatively small-scale sorted bedforms for computational efficiency but results from these experiments are likely applicable to larger scale features (sorted bedforms may be kilometers in scale).

To examine the role of long-term aggradation on the feedback responsible for pattern formation we impose aggradation rates in some experiments of 0.01 m/yr, broken down as 0.003 m/yr coarse sediment and 0.007 m/yr fine sediment (to mimic the model seabed bulk composition). Erosional experiments have an imposed fine sediment erosion rate of 0.003 m/yr. The results of experiments presented here are similar for lower sedimentation/erosion rates.

1.3 Results and Interpretation

1.3.1 Results and Interpretation: ‘Basic’ Scenario; Steady Forcing

The compound interactions in this numerical model and the highly nonlinear, discontinuous nature of the empirical ripple-prediction scheme (see auxiliary information) preclude an analytical explanation of what causes the modes of pattern evolution we observe. Descriptions of dynamics are supported by inspection of the model output and analysis of the equations used as the basis for the modeling effort.

We first consider the case with no storms and no erosion or deposition. Sorted bedforms develop within 20 days (Figure 1 A-B). As a result of initial pattern formation active layer coarseness monotonically decreases from its initial condition (30%) to ~25% (Figure 2A).

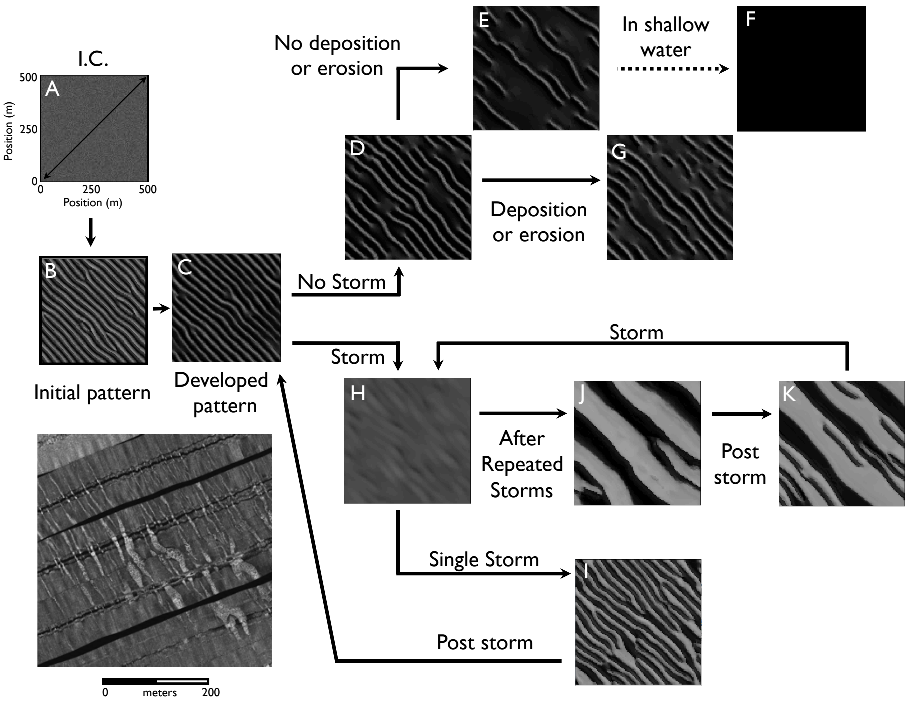

I.C.
A
Position (m)
500
250
0
250
500
Position (m)
B
C
Initial pattern
Developed pattern
No deposition or erosion
E
In shallow water
F
D
G
Deposition or erosion
No Storm
Storm
Storm
H
J
K
After Repeated Storms
Post storm
Single Storm
I
Post storm
0
meters
200

Figure 1: Flow chart of short term sorted bedform pattern evolution. The large lower left panel is an example of bedforms off the coast of Tairua Beach, New Zealand [Coco et al., 2007a] observed in water depths around 5 m. Shoreline is towards the bottom of the panel. Sorted bedforms observed at the same site but in deeper water did not display any significant change over the past 5-10 years (Bryan, pers. comm.). (A) Is the initial condition of the 'Basic' scenario. Black and white pixels indicate fine and coarse sediment respectively. Current direction is from upper right to lower left and reverses daily. (B) After initial pattern development (~200 days). A lack of pattern imperfections (defects) in panel (C) indicates a highly developed pattern (after ~800 model days). These well organized patterns contrast sharply with (D), Pattern intermittence after the active layer reaches the critical state (~2000 days). With no sediment source or sink (E), the pattern is persistent but spatially intermittent. Under high wave conditions or in shallow water the pattern may be ephemeral (F). Deposition or erosion (G) enables a temporally stable pattern with spatial intermittence. Individual coarse domains may be ephemeral. Single storm events (H) tend to destroy bedform relief, mobilize fine and coarse sediment, and coarsen the active layer as fine sediment is brought into suspension. (I) After a storm event the pattern is similar in orientation but coarse domains are wider. Post storm vertical sorting and defect dynamics cause the pattern to evolve into the developed pattern (C). Repeated storm events tend to cause the active layer to coarsen dramatically. Storm events develop large wavelength patterns (J). During inter-storm times (K) this pattern devolves into a superposition of the large wavelength (storm) pattern and the short wavelength (fair weather) pattern. Repeated storms result in bedforms that transit the H-J-K loop, while a cessation of storms cause pattern K to devolve to pattern I (and further C).

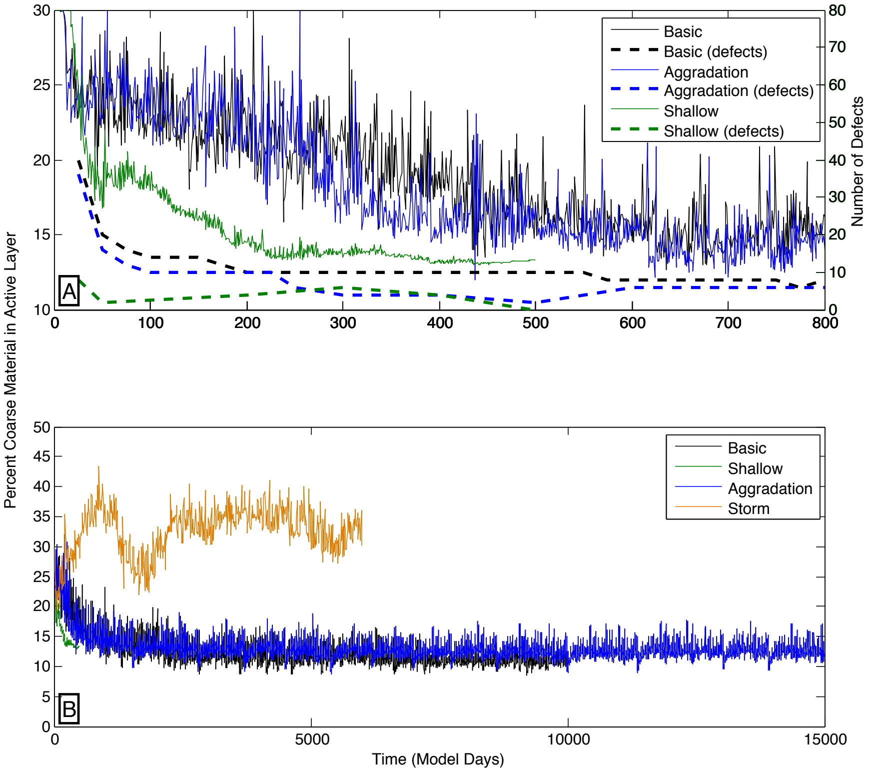


Figure 2: The percentage of coarse material in the active layer and number of defects through time (note the two different time scales). Sorting causes initial fining in the first 20 days of all experiments (A). Defect dynamics (i.e. defect annihilation) and vertical sorting cause subsequent fining. This is most pronounced from day ~ 250 to day ~ 700. A critical level of coarseness is reached in the active layer for most experiments and is maintained for the duration of the experiments (B). The experiment with 'storms' (orange) shows variation in coarseness as a result of autogenic burial and excavation of coarse material.

Active layer fining is accomplished by vertical sediment sorting, a result of changes in current magnitude and direction. Increases in current magnitude give rise to increases in the bathymetric relief of sorted bedforms and vice versa (Coco et al., 2007b). Increases in relief cause fine material in the subsurface to become available for sediment-flow interactions. In this way fine sediment is mined from the seabed below the sorted bedform field. Decreases in relief tend to bury coarse sediment as coarse domains are located in the trough or along the flank of a bedform. Furthermore, daily current reversals and the evolution of the pattern (e.g. wavelength coarsening) cause the coarse domains to migrate. Migration causes the trough of the bedform, the zone previously mined of fine material and a reservoir of coarse material, to change in position. This inhibits the re-excavation of coarse material when subsequent current magnitude increases result in increased sorted bedform relief. In this way the area under the bedform field functions as a sink for coarse material and a source for fine material.

This vertical sorting is especially prevalent during the maturation of the pattern by defect dynamics. In reversing currents, pattern defects tend to migrate farther than the central part of individual domains because the ends of coarse domains contain less coarse material to be transported. During repeated current reversals these fast moving defects tend to continually attach and detach from adjacent coarse domains. Defects are more prone to burial because of these frequent migrations. Unlike defects in bedform patterns undergoing unidirectional flow (Werner and Kocurek, 1999), these modeled defects do not propagate downstream (perpendicular to crests) through the bedform field but instead migrate laterally (parallel to crests). As each defect migrates and the domain it is a part of is progressively buried, the active layer coarseness decreases as does the plan view coverage of coarse sediment. This precipitates an increase in the pattern wavelength to maintain equally spaced bedforms (Figure 1C; Werner and

Kocurek, 1999; Huntley et al., 2008). Surface coarseness decreases from 24% to 16% as the pattern transitions from 14 defects to 8 defects (Figure 2A, between 100 and 600 days). Fining of the active layer as a result of defect dynamics and associated vertical sorting continues until the bedform field reaches a critical level of coarseness (Figure 2).

Active layer fining has consequences for plan view pattern stability. Coco et al. (2007a) found that the average bed composition is an important factor for sorted bedform height and wave ripple characteristics (amplitude and wavelength). Less coarse sediment on the bed results in lower wave ripple amplitude, smaller wave ripple wavelength, lower sorted bedform height and therefore lower bed slope (Coco et al., 2007a). Decreases in ripple size (wavelength and amplitude) bring about a decrease in turbulent intensity above ripples and a decrease in the ability of coarse domains to inhibit fine sediment deposition. Fine sediment is therefore able to dilute coarse domains. Furthermore, lower bathymetric slope as a consequence of lower sorted bedform heights result in higher fluxes of coarse and fine sediment (Coco et al., 2007a). Coarse domains are usually located in the trough or on the up-drift flank of sorted bedforms. This bathymetric slope limits upslope sediment flux (Coco et al., 2007a; Murray and Thieler, 2004). The reduction of this slope results in more mobile coarse sediment that tends to spread from previously coherent coarse domains. The spreading of coarse material from coarse domains and the infiltration of fine sediment tend to shut down the fundamental feedback driving the sorting process.

In experiments with no aggradation/erosion, and with no high wave events, many coarse domains disappear because of this cessation of the sorting feedback. In plan view this process of pattern breakdown occurs on many adjacent coarse patches simultaneously. This leads to a spatially intermittent pattern within 2000 model days, seen in Figure 1E (resulting from the progression C-D-E). Loss of these coarse domains does not ultimately lead to the loss of the entire bedform field, but instead the bedform field maintains a spatially intermittent pattern. This intermittent pattern reflects the critical level of coarseness, an attracting state of active layer composition (Figure 2). In this state there is only enough coarse material on the surface to drive sorting in spatially discontinuous coarse domains.

Complete disappearance of the sorted bedform pattern, a transition from Figure 1E to 1F does not occur under these conditions. Disappearance of sorted bedforms does occur in experiments with shallow water (10 m) and experiments in slightly more energetic wave climates (wave heights ≥ 3 m). Complete disappearance of the pattern in these conditions apparently occurs because stronger wave orbital motions result in more mobile sediment (both coarse and fine) and bedform relief is initially slightly lower (leading to lower bedform slope), which provides less of an impediment to sediment mobility. Furthermore initial sorting buries more coarse material (Figure 2A). The active layer in shallow water experiments falls to a lower critical value of active layer coarseness (13%) much faster (250 days) than experiments in deeper water (16% within

700 days; Figure 2A). This leads to the complete loss of coarse domains within ~400 model days.

1.3.2 Results and Interpretation: Deposition and Erosion

Including erosion or deposition changes the morphology and temporal persistence of sorted bedforms. We performed an experiment for ~15,000 days with steady aggradation and current magnitude adjustments. A sorted bedform pattern persists throughout the entire duration of the numerical experiment, much longer than other reported uses of the model (Coco et al., 2007a, 2007b; Huntley et al., 2008). Bulk deposition during this time amounts to 0.5 m.

Similar to the experiment described above with no deposition, the active layer reaches a critical level of coarseness by processes of vertical sorting. However, with deposition, the continual addition of coarse material at the surface tends to cause coarse domains that were once shrinking to maintain their size or grow. Furthermore, coarse deposition also enables the nucleation of coarse domains where the feedback had originally been shut down. Concurrent fine sediment deposition tends to damp and/or shut down the sorting feedback on several coarse domains. The combined effect of this bimodal deposition is that as soon as the active layer reaches the critical value of coarse material, there is a near constant growth and decay of sorted bedforms (with an identical orientation to the initial pattern). Although the individual coarse and fine domains are not preserved, a pattern dominated by discontinuous coarse domains and many new defects (Figure 1G) is present for the duration of the experiment. The processes that lead to this spatially and temporally intermittent pattern are autogenic, the result of internal dynamics and not a result of spatially or temporally heterogeneous forcing conditions. Experiments with net erosion of fine material yield qualitatively similar plan view patterns and dynamics: the growth and modification of coarse domains occur as a result of the removal of fine sediment.

1.3.3 Results and Interpretation: 'Storm' Scenarios

Pattern dynamics in experiments subject to high wave events (representing storms) put forward another end member for bedform pattern evolution. The original pattern developed under conditions of 2 m high waves is destroyed by the high waves that flatten the sorted bedforms. The active layer becomes more coarse during storm events as fine material is brought into suspension. Coarse material becomes more mobile, spreading over much of the domain because of higher wave forcing and lower bathymetric slopes. As a result mean coarseness of the active layer increases and more coarse material is present over a larger aerial extent of the domain (Figure 1H). Several model days are required after the storm event for a new pattern to emerge that is identical in orientation and similar in wavelength and position to the pre-storm pattern (Figure 1I). Excess coarse material present on the seabed prohibits fine sediment from settling in much of the domain, segregating fine material deposition to distinct areas. This leads to coarse domains that are wider than pre-storm patterns (Figure 1 I vs. C).

Previous work showed that large waves produce sorted bedform patterns with large wavelengths (Coco et al., 2007a). After the high wave event defects begin to migrate, coarse material is buried and fine material begins to infiltrate along the edges of coarse domains. These processes lead to vertical sediment sorting and result in fining of the active layer. Eventually the post-storm arrangement with wide coarse domains reverts to the initial pre-storm pattern (Figure 1C). Without subsequent storms this sorted bedform field will develop as if the storm had never occurred (i.e. the aforementioned 'basic' scenario).

Frequent storm events prohibit the process of pattern maturation (i.e. vertical sorting). This leads to a coarse active layer that exerts a significant control on the plan view pattern. With 100 day storm intervals, coarse domains coalesce into a large wavelength pattern after several storm cycles (Figure 1J). Inter-storm periods characterized by weaker wave climates tend to cause the sorted bedform field to revert to a smaller wavelength pattern. This process of deconvolution gives rise to a sorted bedform pattern that is a superposition of both modes during inter-storm periods (Figure 1K). If storms are frequent enough, sorted bedform fields traverse a repeated loop of J-K-H in Figure 1. If storms are infrequent, Figure 1K eventually devolves to Figure 1I as the active layer fines and the pattern devolves entirely into a small wavelength pattern as coarse material is buried. In storm experiments the active layer does not settle on a critical value for coarseness (Figure 2B). Autogenic fluctuations in coarseness occur around a mean of ~35% coarse material with a 5% amplitude. This variation occurs as a result of autogenic cycles of burial and excavation of coarse material.

1.4 Discussion

Experiments from this study demonstrate that sorted bedform patterns may be persistent or ephemeral based solely on autogenic factors. Ephemeral sorted bedforms, such as those observed by Hume et al. (2003), occur in shallow water where model sorted bedforms have the tendency to be ephemeral. The sorting feedback may be more susceptible to stoppage as a result of the increased active layer fining and vertical sorting in shallow water bedforms or those subject to strong wave forcing. Processes that bury coarse sediment and may impact the strength of the sorting feedback have been observed previously in the Tairua embayment of New Zealand. Green et al. (2004) observed fine sediment infiltration of coarse domain edges and buried fringes of coarse domains are present in short cores (Trembanis and Hume, 2010). The subsequent reappearance of ephemeral bedforms may be a result of intense storms that excavate the buried coarse domains or because of erosion of overlying sediment.

Erosion of fine material or deposition of coarse material allow for the maintenance and nucleation of coarse patches. This is in agreement with the ideas of Green et al. (2004) that dispersed coarse material may be able to seed new bedforms. Erosion or deposition leads to a persistent yet spatially intermittent bedform pattern (as in the basic scenario), but in this situation individual coarse domains may be temporally intermittent features.

Frequent high wave events such as storms tend to coarsen the seabed and contribute to sorted bedform persistence. The sorting feedback is strengthened through storm events that tend to clear or flush coarse domains (at least the edges) of fine sediment, as previously suggested by Green et al. (2004). Frequent storms lead to large temporally stable features with significant heterogeneity in the pattern as a result of the superposition of storm and inter-storm patterns. This superposition of modes rather than the result of bedform migration may be the cause of 'wispy edges' (Murray and Thieler, 2004).

Overall, long term sorted bedform pattern behavior can be encapsulated by nondimensional ratios. Persistent or ephemeral bedforms are separated by a threshold value of water depth over wave height. The numerical location of this threshold is determined by initial and forcing conditions. Another critical parameter is the nondimensional ratio of deposition rate (rate at which sediment is added to the active layer) divided by the rate of active layer fining (rate at which the coarse sediment is removed from the active layer through burial). If this ratio is initially greater than 1, the long term steady state is pattern persistence. If the ratio is initially between 0 and 1 the system organizes itself so that the active layer composition reaches a critical state as seen in the basic and deposition scenarios (Figure 2B). In the limit as this ratio approaches zero, the pattern becomes spatially intermittent but temporally persistent (Figure 1E). Finally, storm behavior can be described by the nondimensional ratio of storm return period divided by the characteristic time scale of active layer fining. In experiments with ratios greater than one (not shown) patterns alternate between storm and non-storm modes (Figure 1 loop C-H-I). Ratios on the order of 1 produce a spatial mixture of storm and non-storm modes (Figure 1 loop H-J-K). For ratios approaching 0 the storm mode will dominate the long term pattern. Future work involving additional experiments could systematically explore and refine the thresholds in these controlling parameters.

Spatial intermittence represents the bistablity of patterned and unpatterned regions, analogous to mixed patterns in Rayleigh-Bénard circulation (e.g. Cross and Hohenberg, 1993). We found only one previously described example of bedform intermittency as a result of autogenic processes, the modeled ripples of Nishimori and Ouchi (1993). Whether of not pattern maturation leads to self-destruction or self-organized intermittency in other bedforms is unknown. Defect dynamics, which tend to drive pattern wavelength increases, can theoretically lead to wavelengths which are too large and consequentially unstable, resulting in pattern reconfiguration through new bedform growth in the trough (Andreotti et al. 2006). If defect dynamics or other nonlinear pattern maturation processes (in the case of sorted bedforms, the burial of coarse material) lead to the inhibition of the fundamental feedback that drives bedform maintenance, other bedforms may be susceptible to this behavior.

1.5 Auxiliary Information

1.5.1 Further details from observed sorted bedforms

Annual surveys show that the sorted bedforms located in the shallow portion (6-16 m) of the Tairua-Pauanui embayment of New Zealand are ephemeral while features in deeper water (18 m to > 30 m) are stable, persistent features [Hume et al., 2003]. In the northern Outer Banks of North Carolina, sorted bedforms found in shallow water (5-6 m) are destroyed by storms but within months tend to reform in similar locations while deeper-water bedforms (5-10 m) seem to be stable under large storm forcing (McNinch, 2004). Deeper sorted bedforms in the North Sea (25-35 m of water) are stable over many decades, with little or no change in position (Diesing et al., 2006). Bedforms located off shore of Martha's Vineyard, MA (7-20 m of water) oscillate in location but have remained in the same position for as long as four decades (Goff et al., 2005). Bedforms in Monterey Bay, CA are mobile in 10-20 m of water but remain stationary at 20-60 m depth (Eittreim et al., 2002). Harrison et al. (2003) describe sand ridges at 3-9 m in depth in west central Florida that become more pronounced after storm events but are destroyed in interstorm time periods (they suggest through bioturbation). Wispy down drift edges of sorted bedforms in the Shinnecock Inlet of Long Island, NY are interpreted as related to sorted bedform migration (Ferrini and Flood, 2005). Murray and Thieler (2004) describe persistent, slowly migrating bedforms off shore (≤ 20 m depth) of Wrightsville Beach, NC with 'wispy' edges that become well defined after large storm events.

1.5.2 Empirical ripple prediction scheme

We use the empirical wave generated ripple scheme from Coco et al. [2007]. Wave generated ripple height (η) and spacing (λ) is determined at each location in the numerical model using the formulation of Wikramanayake and Madsen (1991) combined with the parameterizations of Styles and Glenn (2002):

$$\frac{\eta}{A_w} = \begin{cases} 0.03X^{-0.39} & X \leq 2 \\ 0.45X^{-0.99} & X \geq 2 \end{cases}$$

Equation 1.1

$$\frac{\lambda}{A_w} = \begin{cases} 1.96X^{-0.28} & X \leq 2 \\ 2.71X^{-0.75} & X \geq 2 \end{cases}$$

Equation 1.2

where A_w is the wave orbital amplitude, defined as

$$A_w = \frac{TU_w}{2\pi}$$

Equation 1.3

where T is the wave period and U_w is wave orbital speed. Under conditions of monochromatic waves, the parameter (X) is evaluated as:

$$X = \frac{4\nu {U_w}^2}{d_{50}\left[\left(\frac{\rho_s}{\rho} - 1\right) g d_{50}\right]^{1.5}}$$

Equation 1.4

where ν is the kinematic viscosity, d_{50} is the mean grain size of the bed, g is gravity, ρ is water density and ρ_s is the density of sediment (2.65 gcm^{-3}). Mean grain size at each cell i ($d_{50,i}$) is determined at each time step as:

$$d_{50,i} = (1 - B_{coarse,i})d_{fine} + B_{coarse,i}d_{coarse}$$

Equation 1.5

where $B_{coarse,i}$ is the percentage of coarse sediment in the active layer (the top 0.15 m of the domain) at location i, and d_{fine} and d_{coarse} are the diameter of the fine and coarse fraction, respectively.

2. Prediction of Wave Ripple Characteristics Using Genetic Programming

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

2.1 Introduction

Sufficiently strong water wave propagation over a moveable bed composed of sand grains results in the development rhythmic bedforms whose crest spacing is of the order of cm to m while heights are of the order of cms. These features are often termed vortex ripples because of a recirculation cell that develops on the lee side of the bedform that is subsequently ejected upward during reversals in flow direction. Accurate prediction of vortex ripple size and shape is crucial for successful determination of seabed bottom roughness, a first order control on wave attenuation (e.g., Ardhuin et al., 2002), as well as sediment transport as suspended load (e.g., Green and Black, 1999; Bolaños et al., 2012). Furthermore ripple migration is a fundamental mechanism of bedload transport (e.g., Traykovski et al 1999; Becker et al., 2007), and parameterizations of bedload flux necessitate an accurate depiction of ripple size and shape.

Many predictors of equilibrium ripple geometry have been developed from field and laboratory datasets (e.g. Clifton, 1976; Nielsen, 1981; Grant and Madsen, 1982; Wiberg and Harris, 1994; Faraci and Foti, 2002; Styles and Glenn, 2002; Grasmeijer and Kleinhans, 2004; Soulsby and Whitehouse, 2005, 2012; Pedocchi and García, 2009a; Camenen, 2009). Equilibrium ripple size and shape is frequently broken down to include 3 subpopulations, a convention developed by Clifton (1976), and reviewed here in order of increasing hydrodynamic forcing. Orbital ripples are believed to scale linearly with wave orbital diameter at the seabed and display the largest steepness (ripple height/wavelength ~ 0.15). Suborbital ripples show spacing that depends on wave orbital diameter and grain size. In even stronger hydrodynamic conditions anorbital ripples form, whose size is related to grain size alone and whose scaling is irrespective of wave orbital diameter. Suborbital ripples link the population of anorbital ripples with those of orbital ripples.

As noted by Smith and Wiberg (2006), recent field and laboratory work has challenged the existing typology for wave-generated ripples as a result of the addition of two new populations (Figure 3).

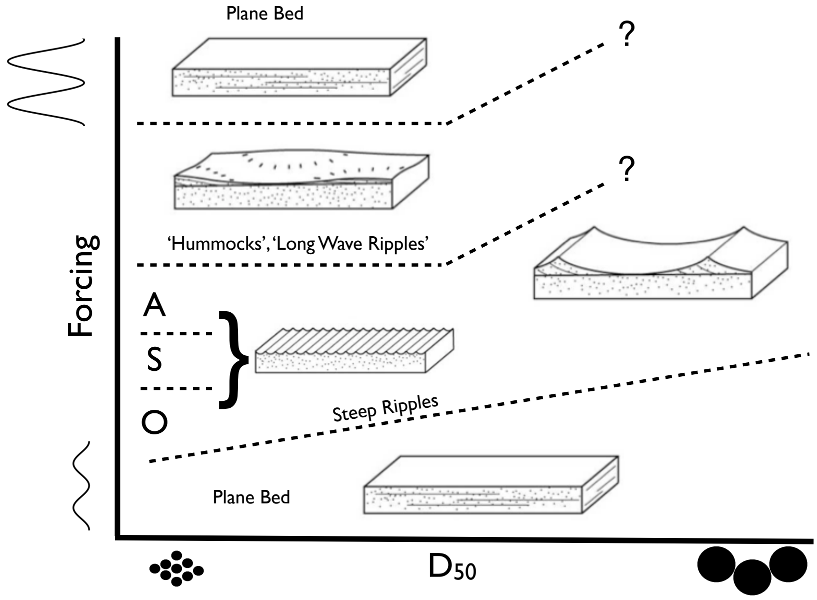


Figure 3: A schematic phase diagram of oscillatory bedforms. O, S, and A represent orbital, suborbital, and anorbital ripples respectively: smaller steep ripples that occur under small/moderate hydrodynamic forcing in fine sands. Orbital, suborbital and Anorbital ripples occur in sequence as hydrodynamic forcing is increased. Recent data collection campaigns have focused on 1) strong hydrodynamic forcing in fine sands ('hummocks' or 'long wave ripples') and 2) steep, large ripples in coarse sand. Modified after Cummings et al., (2009). Question marks denote the unknown threshold for plane bed in coarse grained environments, and unknown potential for coarse grained environments to be sculpted into long wavelength 'hummocky' ripples. Additionally it is unknown if suborbital and anorbital scale ripples exist in coarse grain settings. Lower plane bed conditions are likely only applicable for laboratory studies where the bed is artificially flattened (field conditions retain relict or antecedent bed geometry).

The first are ripples measured in fine sand under strong hydrodynamic conditions. Field and laboratory campaigns in more energetic conditions have discovered the presence of long wavelength, low amplitude ripples ('hummocks') in fine sands that scale with orbital diameter (e.g. Hanes et al., 2001, O'Donoghue et al., 2006). Predictors are unable to accurately capture this ripple size and shape (e.g. Bolaños et al., 2012), yet modeling (Chang and Hanes, 2004) and observation (Green and Black, 1999; Cummings et al., 2009) of these bedforms show they eject vortices and are therefore important for their influence on seabed roughness and sediment transport. Furthermore at times these long wavelength ripples have superimposed anorbital ripples (e.g. Southard et al., 1990; Hanes et al., 2001; Williams et al., 2004), another unsolved problem in wave ripple prediction. Because of these complications, Pedocchi and García (2009a), who developed a recent well performing predictor, omit long wavelength ripples from their analysis, but note that these long wavelength 'round crested' ripples are observed above a critical threshold in U/w_s (where U is the maximum orbital velocity at the bed and w_s is the sediment fall velocity). Dumas et al. (2005) and Cummings et al. (2009) also show that the transition from anorbital scale ripples to round crested long wave orbital scale ripples is a function of orbital velocity (a set value for their given sediment mixtures).

The second new population of ripples are those found in medium to coarse sand (Traykovski et al., 1999; Ardhuin et al., 2002; Becker et al., 2007; Masselinsk et al., 2007; Traykovski, 2007; Cummings et al., 2009; Yamaguchi and Sekiguchi, 2011). Coarse grained ripples have been observed in shelf environments for several decades (e.g. Forbes and Boyd, 1987; Leckie et al., 1988 and references therein) but until recently ripple measurements have not been coupled to the hydrodynamic parameters of their formation. Recent lab work by Cummings et al., (2009) demonstrated the persistence of steep ripples with orbital scaling in coarse sand under strong hydrodynamic conditions.

These two new populations of ripples highlight a perennial problem with empirical predictors; unless equations are built using large, integrated data sets that encompass many conditions, prediction schemes are difficult to translate to different settings. A non-empirical approach, such as models based on first principles (e.g. Foti and Blondeaux, 1995; Blondeaux, 2001; Charru and Hinch, 2006), presents different problems: nonlinear, emergent processes that occur at the ripple scale such as flow separation, vortex ejection, turbulence, sediment suspension, pattern coarsening, defect creation, migration and annihilation (Werner and Kocurek, 1999), and the existence of multiple stable configurations in ripple sizes/shapes at a given hydrodynamic condition (a stability balloon; Hansen et al., 2001) limit the usefulness of finite-amplitude predictions. Prediction by numerical models of coupled fluid flow and bed evolution present promising results but have so far been tested under a narrow range of conditions and compared to few data sets (Marieu et al., 2008; Chou and Fringer, 2010).

If empirical data driven predictors are currently the most broadly applicable tools to develop field scale predictions, how should they be built? Traditionally the development of an empirical predictor relies on transforming a single (or several) noisy multidimensional dataset to lower-dimensions and fitting a curve (with a set functional form) through the resultant point cloud. Here we offer a different solution: a data integration campaign (the collection of many published datasets) followed by machine learning (ML), whereby computational optimization techniques are used to find solutions to multidimensional and nonlinear problems. The suite of techniques encompassed by ML are essentially identical to empirical data driven techniques used previously except the trial and optimization of solutions is outsourced to a computer.

The most common ML paradigm used in coastal studies is artificial neural networks (ANN). Recent examples of its use include predictions of alongshore sediment transport in the surfzone (van Maanen et al., 2010), sand bar behavior (Pape et al., 2010) and suspended sediment reference concentration under waves (Oehler et al., 2012). Yan et al. (2008) used an artificial neural network to predict wave ripple geometry (length and height) based on three input parameters (median grain size, wave period, and the maximum near bed wave orbital velocity). ANN results give better predictions based on 3 statistical measures (scatter index, correlation coefficient, and mean geometric deviation) than four common empirical models (Nielsen, 1981; Van Rijn, 1993; Wiberg and Harris, 1994; Grasmeijer and Kleinhans, 2004). Yet the ANN ripple prediction scheme derived by Yan et al. (2008) was developed and compared to a limited dataset. Furthermore ANNs are problematic because the highly nonlinear result is difficult to interpret and does not offer immediate insight into the physical nature of the problem at hand. Decision or regression trees (e.g., Oehler et al., 2012), another common and well performing ML technique, is also hampered by the lack of direct physical significance and other drawbacks such as the lack of smoothness.

In this contribution we use genetic programming (GP; Koza, 1992), a population based optimization technique where the population consists of individual equations (i.e. a population of individual predictors). The mathematical or logical operations that constitute each algorithms can be modified at every time step via an 'evolutionary' process (such as crossover and mutation) to produce expressions that optimize model-data fit. Outputs developed by GP can be smooth functions that are easy to examine and interpret for physical significance. Furthermore, *a priori* determination of the functional form of the predictor is not required and the final optimized solution can take on any mathematical form (within user defined limits). Thus far genetic programming has been applied to a wide range of problems including the prediction of freshwater phytoplankton dynamics (Whigam and Recknagel, 1999), downscaling of atmospheric model output (Coulibaly, 2004), determining appropriate parameterization for roughness in vegetated flows (Baptist et al., 2007), wave forecasting (Kambekar and Deo, 2012) and mapping of seafloor habitats (Silva and Tseng, 2008).

The goal of this manuscript is to demonstrate the applicability of ML techniques (specifically GP) to research questions in the coastal domain. To accomplish this goal we compile 27 different field and laboratory data sets of wave ripple prediction (995 individual measurements; Table 1) that span a broad range of conditions and develop a new wave-ripple predictor that is able to capture the morphology of ripple geometry in a wide range of forcing conditions, including conditions where long wave orbital ripples are present. We put our results in the context of existing formulations and theories, and assess the physical relevance of GP predictors. Our new equilibrium predictor ignores the effect of ripple orientation, time evolution, heterogeneous sediment, superimposed current, ripple asymmetry, and bio-degradation of ripples. We discuss these limitations in the discussion section but note here that other existing time dependent ripple prediction schemes capture one or more (but not all) of these processes (i.e., Soulsby et al., 2012; Traykovski 2007). Finally, the compilation of published ripple data allows for the identification of gaps in knowledge and observations that should be pursued in future research. Future data collection campaigns can be added to this database, allowing for modifications to the prediction schemes shown below. In this sense the ripple prediction scheme we demonstrate here is dynamic.

2.2 Data

As a result of decades of study, many wave ripple datasets are available in the scientific literature. Examples of recent wave ripple data integration and compilations are Soulsby and Whitehouse (2005), Pedocchi and García (2009a) and Camenen (2009). Here we follow the lead of Pedocchi and García (2009a) and limit our data collection to studies using sediment with quartz (or near quartz) densities (2.65 g/cm^3) performed in large oscillatory tunnels, large wave flumes, wave racetracks and field conditions (i.e. we omit oscillating trays). Data on rolling-grain ripples, small bedforms that initially appear when flat beds are subject to oscillatory water motion, are ignored in this study because they have been experimentally shown to be a transient stage of ripple evolution (Faraci and Foti, 2001). We use 27 published studies in our dataset. Each measurement contains wave ripple, hydrodynamic, and sedimentological parameters (Table 1).

Table 1: Data Summary; Measurement numbers reported are the ripple length measurements used in our study. Measurements with both length and height are less.

Authors	Setting	Measurements	T (s)	U (m/s)	d_0 (m)	D_{50} (m)
Boyd et al. 1988	Field	36	3.8-9.8	0.04-0.28	0.05-0.60	0.00011
Cummings et al. 2009	Wave Racetrack	14	4.4-14	0.37-1.22	0.95-4.50	0.00012-0.0008
Delgado Blanco et al. 2004	Wave Flume	17	6.0	0.14-0.74	0.27-1.42	0.00035
Doucette 2000	Field	49	4.7-12.2	0.15-0.52	0.31-1.93	0.00015-0.00053
Doucette 2002	Field	25	2.2-12.2	0.17-0.66	0.31-2.22	0.00035-0.00062
Doucette and O'Donoghue 2006	Osc. Tunnel	32	2.0-12.2	0.29-0.63	0.24-2.00	0.00044
Dumas et al. 2005	Osc. Tunnel	23	7.9-11.0	0.21-1.26	0.51-4.17	0.00011-0.00023
Grasmeijer and Kleinhans 2004	Field	26	4.0-10.5	0.23-0.84	0.58-2.41	0.00024
Hanes et al. 2001	Field	169	7.1-19.7	0.92-1.11	0.47-5.02	0.00012-0.00166
Hume et al. 1999	Field	9	11.0	0.08-0.37	0.30-1.30	0.00040
Inman 1957	Field	59	0.5-15.0	0.06-0.94	0.04-2.74	0.00008-0.00091
Kennedy and Falcon 1965	Wave Flume	10	1.1-2.0	0.12-0.26	0.04-0.13	0.00010-0.00032
Miller and Komar 1980a	Wave Flume	4	3.0-8.0	0.05-0.34	0.14-0.54	0.00017
Miller and Komar 1980b	Field	26	6.0-18.2	0.03-0.41	0.07-2.14	0.00017-0.00029
Mogridge 1972	Osc. Tunnel/ W. Flume	72	1.0-14.0	0.13-0.68	0.05-1.84	0.00036
O'Donoghue and Clubb 2001	Osc. Tunnel	35	2.0-15.0	0.25-0.94	0.16-2.92	0.00018-0.00044
O'Donoghue et al. 2006	Osc. Tunnel	27	3.1-12.5	0.31-0.85	0.42-2.70	0.00022-0.00044
Pedocchi and García 2009b	Osc. Tunnel	22	2.0-18.0	0.20-1.00	0.16-2.86	0.00025
Ribberink and Al-Salem 1994	Osc. Tunnel	25	2.0-10.0	0.30-1.50	0.31-3.82	0.00021
Sleath 1982	Osc. Tunnel	13	2.9-5.1	0.16-0.44	0.17-0.51	0.00020-0.00041
Sleath and Wallbridge 2002	Osc. Tunnel	26	2.8-5.	0.08-0.77	0.12-0.80	0.00020-0.00080
Southard et al. 1990	Osc. Tunnel	63	93.1-19.3	0.16-1.00	0.26-3.56	0.00011-0.00032
Thorne et al. 2002	Wave Flume	14	4.0-6.0	0.26-0.66	0.41-1.05	0.00033
Williams et al. 2000	Wave Flume	9	4.8-5.3	0.19-0.69	0.30-1.10	0.00016-0.00033
Williams et al. 2004	Wave Flume	65	4.0-6.0	0.13-1.02	0.25-1.96	0.00016-0.00035
Xu 2005	Field	13	8.9-14.8	0.11-0.16	0.41-0.76	0.00009
Yamaguchi and Sekiguchi 2011	Wave Flume	111	1.3-5.0	0.18-0.51	0.07-0.55	0.00032-0.00073

The dataset is split 59% /41% between laboratory and field conditions, and laboratory measurements are obtained from a 49% /49%/2% split between oscillatory tunnels, wave flumes, and wave racetracks. Measurement error is different for each data set in our database, a natural consequence of data integration campaigns that assemble data collected by different instruments and techniques. We assume that measurements of ripple data obtained in field settings are at or near equilibrium.

Our database can be visualized as a series of histograms showing the parameter range in our dataset (Figure 4).

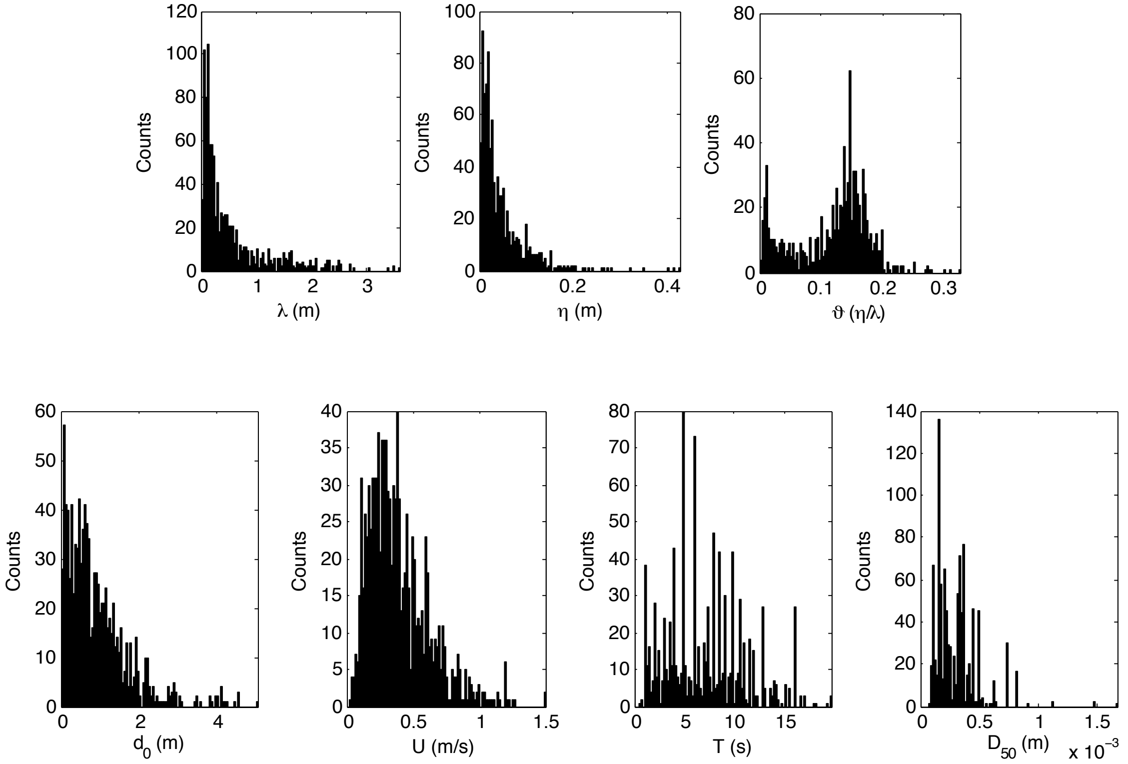


Figure 4: Histograms for ripple length (995 measurements), ripple height (872 measurements), ripple steepness (872 measurements), and for hydrodynamic and sedimentological variables used in this study (includes all 995 data points). Note the different Y-axis values for each graph.

A majority of ripple measurements in our database occur at hydrodynamic conditions of $d_0 < 2$ m, $U < 0.75$ m/s and sedimentological conditions of $D_{50} < 0.5$ mm. Another notable attribute is the strong bimodal signature of ripple steepness centered at values of ~0.15 and ~0.01. These clusters represent steep ripples and 'hummocky' ripples, respectively. We base our prediction of wave ripple wavelength λ (m), ripple height η (m), and ripple steepness ϑ (η/λ; dimensionless) on four variables: wave period T (s), bottom orbital excursion d_0 (m), median grain size D_{50} (m), and maximum near bed orbital velocity U (m/s). A hallmark of field data sets is the irregular forcing, requiring us to reconcile different measured parameters. Several field datasets used in the compiled dataset reported hydrodynamic parameters in terms of significant values (U_{sig}, $d_{0,sig}$, and T_{sig}). We followed the protocol of Pedocchi and García (2009) and assume $U = U_{sig}$ (and furthermore $d_0 = d_{0,sig}$ and $T=T_{sig}$). We acknowledge that the merging of disparate data sources introduces uncertainty into the data.

The hydrodynamic and sedimentological conditions covered by this dataset can be visualized using 6 projections of the 4 dimensional phase space (Figure 5).

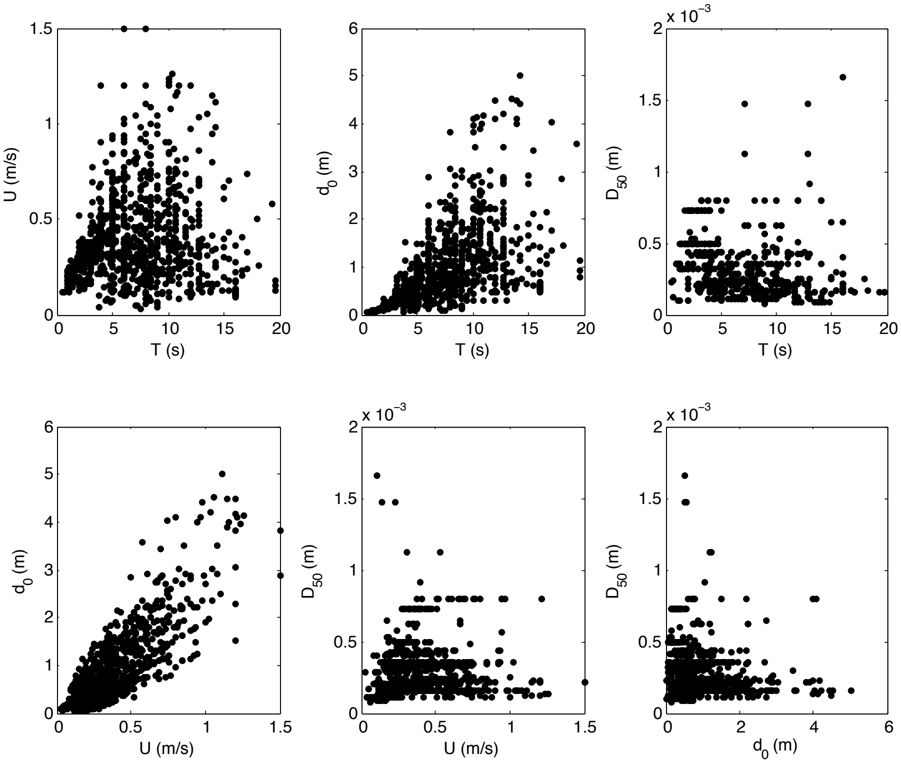


Figure 5: Visualization of the range of forcing conditions in the ripple length dataset. Each plot represents a 2 dimensional projection of the entire data set onto the set of axes shown. For instance, the first panel with data projected onto the U-T plane shows no information about D_{50} or d_0. Ripple height dataset shows qualitatively similar distribution and range, but with fewer data points (872 vs. 995).

Notable sparseness occurs in this database at strong hydrodynamic conditions, and at median seabed grain sizes above 0.5 mm. We use T, d_0, and U as separate independent variables for input to the GP (though they are related by $d_0=UT/\pi$) in an attempt to introduce no additional information about which of these parameters is most relevant. As GP is a data driven technique, the raw hydrodynamic data is given as input and the ML process determines which hydrodynamic variable(s) is most relevant from a statistical standpoint. We use T, d_0, D_{50}, and U to predict λ. Predicted λ is incorporated into the suite of variables (i.e., T, d_0, D_{50}, U) used to predict η. We combining the predictors for λ and η enable the development of a predictor for ripple steepness.

Yet we do not enforce the accurate depiction of steepness in the development of ripple height and length predictors and imprecision in the λ and η equations may cause imprecision in the prediction of ϑ. However, in some circumstances accurate depiction of height and steepness is required for the parameterization of relevant processes (e.g. vertical suspended sediment diffusivity; Nielsen, 1992): therefore we also develop an independent ripple steepness predictor using the genetic programming technique. The development of a third predictor also further demonstrates the strengths and weaknesses of GP and ML techniques. Predicted λ and η are added to the variables (T, d_0, D_{50}, U) used to predict ϑ. The development of predictors for λ, η and ϑ without enforcing interoperability relies on users to decide which predictors are most important for the specific research question.

Several published studies measure two superimposed ripple scales (larger orbital scale ripples and smaller anorbital scale ripples) at a single hydrodynamic condition (e.g. Hanes et al., 2001; Pedocchi and García 2009b; Cummings et al., 2009). Work by Cummings et al., (2009) shows that both pattern modes occur as maximum orbital velocity is increased and the ripple pattern transitions from small scale (anorbital) ripples to large scale orbital ripples (‘hummocks’). Upon further velocity increase, the small scale ripples are destroyed and only the large scale orbital features remain (Cummings et al., 2009). The threshold of large scale orbital ripple appearance can be estimated from the work of Pedocchi and García (2009a) who found that large scale features appear at a threshold value of $U/w_s \cong 25$. When both anorbital and large scale ripples are present in tabulated data (e.g. Hanes et al 2001) we only include large scale ripples: the scaling of long wave ripples with bottom orbital diameter suggests a physical relationship to small scale orbital and suborbital ripples. In contrast, anorbital ripples scale with grain size (similar to current ripples) and the mechanism responsible for their formation may be different

(Wiberg and Harris, 1994). We remove small-scale (anorbital) ripples from our database if they are present at values of $U_0/w_s \geq 25$; laboratory work by Cummings et al., (2009) and Pedocchi and García (2009a) suggests that this regime is dominated by large scale ripples. The targeted collection of field and laboratory data is needed to refine this threshold.

2.3 Methods

2.3.1 Selection of training, Validation, and Testing Data

The database is split into three subsets to be used as training, validation, and testing. The GP algorithm uses the training dataset to develop and optimize candidate solutions. The validation dataset is used to evaluate the fitness of GP derived solutions and define which predictors persist. Testing data is not used or seen by the GP algorithm and is instead reserved as an independent test of the final predictors (and other published predictors). In the genetic programming literature there remains no proven 'best practice' for percentage of training, validation, and testing data, nor a well defined method of splitting these datasets. This may be because data splitting (e.g., the retention of a testing dataset) is not addressed in the foundational literature of the technique (as noted by Kushchu, 2002). Yet because our database of ripple measurements contains only sparse data at energetic hydrodynamic conditions and large grain sizes, the selection and partitioning of data into these three categories is crucial to develop a well performing predictor applicable to a range of environments (Bowden et al., 2002). For example, random division of the data has the potential to produce a significant problem; the training data is likely to misrepresent the full phase space of the entire dataset (i.e. exclude coarse grained and/or strong hydrodynamic data).

Informed data selection (i.e. selection based on clustering) 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; e.g. Camus et al., 2011). This algorithm is not a clustering routine (where cluster centroids are selected to represent a representative value of the data in the cluster), but instead a selection routine (where a centroid represents the most dissimilar data point from the previous centroids; Camus et al., 2011). Though our selection technique is different than the clustering techniques used by Bowden et al. (2002), our approach leads to a similar result: the use of a minimum of training data that is able to capture the variance in hydrodynamic and sedimentological conditions of the entire dataset while leaving more data to be used as validation and testing.

Our implemented version of the maximum dissimilarity algorithm is based on the description provided in Camus et al. (2011). Selection starts with the normalization of the data to a value between 0 and 1:

$$X_n = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Equation 2.1

where X_n is the new normalized data value (between 0 and 1), X is the original value, X_{min} and X_{max} are the minimum and maximum of all values of variable X, respectively. After this normalization a single data point, a 'seed', is selected as the first centroid. Since our dataset is typified by sparseness in the coarse grain data, we use the largest grain size measurements as the first centroid (the 'seed'). The user selects the number of centroids and the algorithm then selects the additional centroids through an iterative process: Each data point in our data set is a 4-dimensional vector (normalized T, U, 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). This routine continues until the user defined number of centroids is reached, after which data is denormalized.

There remains significant ambiguity in determining the appropriate number of centroids (or clusters) needed to accurately represent data, especially continuous data (e.g. May et al 2010). Our dataset on wave ripples is multidimensional and relatively continuous (i.e. not naturally clustered). Furthermore the dataset is sparse in areas because of a lack of collected data, while densely populated with measurements in other regions of phase space (e.g., experimental campaigns at specific hydrodynamic and/or sedimentological conditions). Since we intend to use selected centroids as representatives of the entire dataset, selecting too many centroids will likely rob the validation and testing datasets of poorly represented data (e.g., large T, U, d_0, D_{50}) while too few centroids will leave the testing data with to few data to capture the variability in the dataset. We use 30 centroids for the prediction of λ and 40 centroids for the prediction of η. Centroids used to represent η are also used for analysis of ϑ. Centroid locations can be seen in Figure 6.

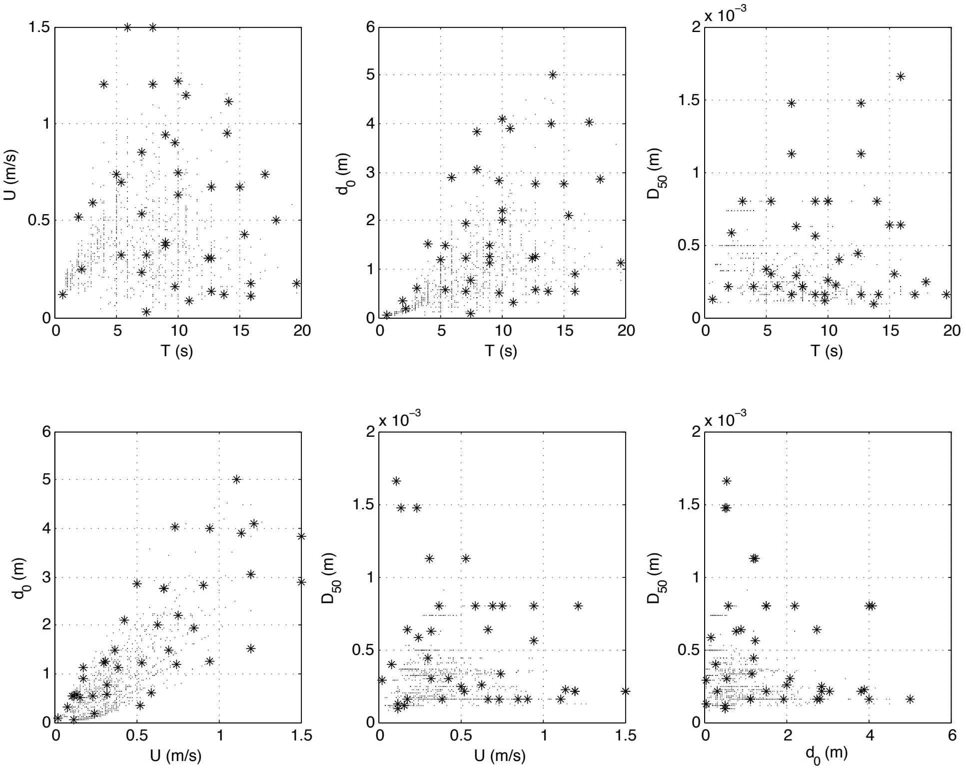


Figure 6: Centroid locations in the ripple length dataset, visualized using the projections shown in Figure 5. Stars denote centroid locations (training data), while points denote unselected data (validation and testing). Note that centroids are distributed throughout the dataset. Centroid locations for the ripple height (and steepness) dataset look qualitatively similar but have more centroids (40 vs. 30) and fewer data points (832 vs. 965).

The use of fewer centroids (10-20) produced too few predictors while more centroids (~100) tended to produce many more nonlinear and potentially overfit solutions. In addition, the solutions obtained with more centroids were qualitatively similar to the solutions presented below using only 30-40 centroids. More centroids are used to predict ripple height because η is more difficult to predict (see also Yan et al., 2008; Williams et al., 2004). This is likely a result of the nonlinearities associated with ripple crests protruding into regions of flow with higher velocities: ripple height is likely more strongly influenced by suspension processes as a result. Data selected as the centroid locations are used for the training data. The points not selected as centroids (i.e. not selected as training data) are used for validation and testing data. Data is split between validation and testing randomly, without using a selection routine. Therefore the breakdown for the λ datasets is ~3% training, ~48% validation, ~48% testing, while the η (and ϑ) dataset breakdown is ~5% training, ~47% validation, ~47% testing.

2.3.2 Genetic Programming

We operate on this compiled ripple data using the evolutionary computation technique of genetic programming (GP), a ML paradigm whereby candidate solutions (in the form of randomly generated equations) are evaluated and subsequently modified (Koza, 1992; Poli et al., 2008). The modification of candidate solutions is manifest as changes in variables and mathematical relationships between variables (i.e. the mathematical form), hence the description of this style of problem as 'symbolic regression'. Variables used in this study to predict wave ripple geometry are T, U, d_0,

D_{50}, λ (for height and steepness prediction), η (for steepness prediction), as well as GP derived constants. Nondimensional, renormalized input (from 0-1) is not necessary with GP (as it is with other ML techniques), and input is fed into the algorithm with units. Only D_{50} is renormalized in this analysis, and fed into the GP in units of mm (as opposed to m, but the presentation of all results in this contribution are in meters). Mathematical operators used in this study are + (addition), − (subtraction), × (multiplication), ÷ (division), √ (square root), as well as integer powers ($e.g.\ \ x^2, x^3, x^4$). Furthermore we omit logical functions (e.g. if-then-else) because of the lack of smoothness when incorporating these components.

Candidate solutions are evaluated based 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:

$$MSE = \frac{(p-b)^2}{n}$$

Equation 2.2

where MSE is the Mean Squared Error, n is the sample size, p are the predicted values, and b are the observed values. The correlation coefficient, one of the error metrics used in previous ripple studies (Yan et al., 2008), was not used as a fitness function because it tended to develop nonphysical predictors (negative wavelengths and heights under certain conditions) that matched the shape of the data but did not align well with actual magnitudes.

Equations that minimize mean squared error are retained, while poor performing solutions are discarded. Retained solutions are combined, rearranged and manipulated in a probabilistic manner according to evolutionary 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. As an example, candidate solutions are commonly encoded in GP software as ‘trees’, and the modification of candidate solutions (change of variables and/or mathematical expression) is accomplished through adjustments in tree ‘limbs’ (Figure 7).

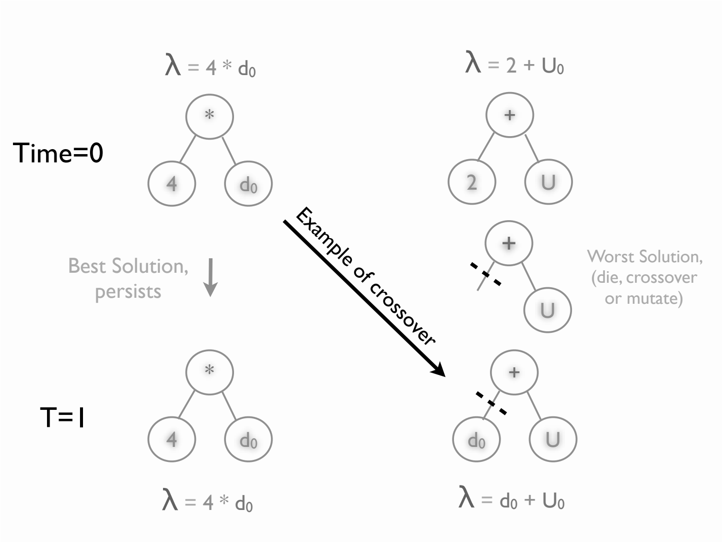


Figure 7: Example of the genetic programming process. Potential solutions are encoded as a population of ‘trees’. Here a hypothetical population of two solutions is shown. The first solution has a low MSE and therefore persists to the next iteration. The second solution has a high MSE and therefore is subject to removal, mutation, or crossover. Here is an example of ‘crossover’ whereby the old solution is combined with parts of other, better performing solutions to create a new potential solution in the next iteration.

Through time predictors gain complexity (i.e. trees grow in size) as they are recombined in a variety of ways, moving from simple equations (e.g. two variables and one mathematical symbol linking them) to highly nonlinear, complex expressions (e.g. many variables linked by many symbols). In this way the growth and adjustment of candidate solutions enables the searching of an increasingly larger phase space (i.e. variable and symbolic space), and find optimized solutions to the problem at hand. This search process occurs until a solution with zero error is found or the routine is terminated.

In this study we use a proven symbolic regression/genetic programming software package developed by Schmidt and Lipson (2009; 2013). This software package, 'Eureqa', modifies the tree-based encoding outlined above by eliminating redundancy when multiple 'tree limbs' are identical. The software output is a suite of solutions with increasing mathematical 'complexity', where complexity is a count of the numbers of operations and variables used in the candidate solution. Each solution of a given complexity represents the equation with the least error compared to identically 'complex' candidate solutions. Furthermore, to be retained in the solution set, a given solutions must have less error compared to all previous less-complex solutions. Therefore the suite of solutions that is developed as output lie along the 'Pareto front', a line in complexity-fitness space that illustrates fitness increases with the increasing complexity of candidate solutions. Because simple predictors are retained though more complex predictors may fit the data with less error, the user must pick a single solution as the final predictor of choice.

2.3.3 Generalization and Overfitting

The lack of a single optimal solution as output from the GP algorithm is likely a consequence of using noisy data (e.g., field data) and examining a phenomena that may not have a single solution, but instead a small range of possible solutions (i.e. there may be multiple stable ripple configurations for a given hydrodynamic/sedimentological condition, a ‘stability balloon’; Hansen et al., 2001). The determination of an ideal solution from the GP program was further complicated because there is no stoppage routine built into the algorithm (e.g., based on fitness) used in this study. We cease the search after roughly 10^{10} formulas have been evaluated as continued search shows only marginal increases in predictive power (and this increase occurs only on more complex, likely overfit, predictors). The solutions were then evaluated to determine the most appropriate final predictor. Several methods for eliminating overfit solutions exist (e.g. Gonçalves et al., 2012). We use several techniques in parallel to determine appropriate solutions: 1) bias toward shorter, physically reasonable solutions, 2) examining ‘cliffs’ in the Pareto front, and 3) examination of solution fit.

Many of the more complex solutions have lower error with training and validation data but are physically uninterpretable. Therefore when evaluating the family of solutions from a given genetic programming iteration we tend to bias our search for the most universal predictor by preferring compact solutions because they tend to offer more generalization and are likely less overfit (The minimum description length principle; e.g. O'Neill et al., 2010). Shorter solutions reappear with repeat initialization of the genetic programming algorithm, suggesting that these represent the globally optimum solutions for a given function size. Longer solutions do not tend to reappear, either 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.

Aside from examining the solutions from least complex to most complex, examining areas along the Pareto front where large gains in prediction are obtained with small gains in solution complexity is a natural place to observe potential solutions (Figure 8).

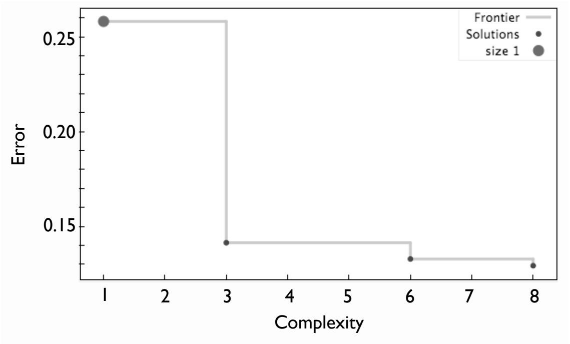


Figure 8: Ripple Length Pareto front; Error is expressed as mean squared error of candidate solution versus the validation data set. Complexity is a quantification of the candidate solution length (both mathematical operators and variables).

These areas along the Pareto front are referred to as ‘cliffs’. Schmidt and Lipson (2009) used the last of such 'cliffs' to observe many physically relevant solutions. In this study final solutions were chosen from the subset of solutions that are ‘cliffs’ along the Pareto front

Candidate solutions are evaluated by minimizing error functions. Occasionally candidate solutions are able to minimize the mean squared error but provide unphysical solutions (e.g. negative ripple wavelengths under some conditions) or generally poor global performance (e.g. flat, constant predictors). These solutions must be manually disregarded, as there is as yet no means of excluding them.

2.3.4 Comparison with other predictors

Predictor performance is evaluated, using the independent testing data, with the Normalized Root Mean Squared Error (NRMSE):

$$NRMSE = \frac{\sqrt{MSE}}{\bar{b}}$$

Equation 2.3

where $\bar{b}$ is the mean of the observed values. Additionally we report correlation coefficient (Pearson’s *r*) for each predictor evaluated against the independent testing data.

We compare our results to two recently developed and widely used predictors: Soulsby and Whitehouse (2005; also reported in Soulsby et al., 2012) and Pedocchi and

García (2009a). As noted by Soulsby et al., (2012), recent work by Camenen (2009) using a large compiled database of ripple measurements found the Soulsby and Whitehouse (2005) formulation to be the best overall predictor compared to those developed by Grant and Madsen (1982), Wikramanayake and Madsen (1991), Van Rijn (1993), Mogridge et al., (1994), Wiberg and Harris (1994), and Grasmeijer and Kleinhans (2004). The recent work of Pedocchi and García (2009a), which was not evaluated by Camenen (2009), yields good collapse of the data compared to other the predictors mentioned above and performs well in field conditions (Bolaños et al., 2012). The Soulsby and Whitehouse (2005) predictor for length and steepness (η/λ) is:

$$\frac{\lambda}{A} = \left[1 + (1.87\times10^{-3})\frac{A}{D_{50}}\left(1 - e^{\left\{-\left(2.0\times10^{-4}\frac{A}{D_{50}}\right)^{1.5}\right\}}\right)\right]^{-1}$$

Equation 2.4

$$\frac{\eta}{\lambda} = 0.15\left[1 - e^{\left\{-\left(5000\frac{D_{50}}{A}\right)^{3.5}\right\}}\right]$$

Equation 2.5

where A is the wave orbital amplitude (2A=d_0). Combining (4) and (5) yields η alone. The Pedocchi and García (2009a) predictor is:

$$\frac{\lambda}{d_0} = \begin{cases} 0.65\left[\left(0.050\frac{U}{w_s}\right)^2 + 1\right]^{-1}, & Re_p \geq 13 \\ 0.65\left[\left(0.040\frac{U}{w_s}\right)^2 + 1\right]^{-1}, & 9 \leq Re_p < 13 \\ 0.65\left[\left(0.054\frac{U}{w_s}\right)^3 + 1\right]^{-1}, & Re_p < 9 \end{cases}$$

Equation 2.6

$$\frac{\eta}{d_0} = \begin{cases} 0.1\left[\left(0.055\frac{U}{w_s}\right)^3 + 1\right]^{-1}, & Re_p \geq 13 \\ 0.1\left[\left(0.055\frac{U}{w_s}\right)^4 + 1\right]^{-1}, & 9 \leq Re_p < 13 \\ 0.1\left[\left(0.055\frac{U}{w_s}\right)^5 + 1\right]^{-1}, & Re_p < 9 \end{cases}$$

Equation 2.7

where w_s is evaluated for D_{50} and Re_p is a dimensionless particle size (Pedocchi and García, 2009a) evaluated as:

$$Re_p = \frac{\sqrt{gRD_{50}}D_{50}}{\nu}$$

Equation 2.8

where g is gravity, R is the submerged specific density of sediment (here taken to be 1.65) and ν is kinematic viscosity. The three size classes ($Re_p \geq 13, 9 \leq Re_p < 13$ and $Re_p < 9$) correspond to coarse, medium and fine sand respectively and the three separate equations result in slight discontinuities.

Lastly we note that we are unable to compare the performance of our GP derived predictor to the ANN model developed by Yan et al., (2008) as we do not know the final optimized ANN equation developed by Yan et al., (2008). In addition, we do not know which data was used as training/validation or testing in the development of the ANN model.

2.4.Results

2.4.1 Ripple Wavelength

The GP algorithm output is shown in Table 2.

Table 2: Solutions for Ripple Length

Solution	Complexity	MSE
$\lambda = U$	1	0.258
$\lambda = 0.607d_0$	3	0.141
$\lambda = \frac{d_0}{1.39 + (1000D_{50})}$	6	0.133
$\lambda = \frac{d_0}{1.12 + 2.18(1000D_{50})}$	8	0.129

This experiment evaluated 10^{10} formulas to develop the Pareto front shown in Figure 8. Cliffs, significant gains in error for small changes in equation complexity occur along the Pareto front at complexities of 3, 6, and 8 (Figure 8) The first of these cliffs (at complexity 3) is a predictor, $\lambda = 0.607d_0$, that mimics the basic form of the orbital scale (i.e. weak hydrodynamics) predictor commonly used today, where ripple wavelength is a linear function of orbital excursion (e.g. $\lambda = 0.65d_0$ from Miller and Komar, 1980a; $\lambda = 0.62d_0$ from Wiberg and Harris, 1994). Debate surrounds the correct value of the coefficient modifying orbital excursion, especially in medium to coarse sand (e.g. Becker et al., 2007; Traykovski et al., 1999). All solutions that are more complex than the solution of complexity 3 demonstrate why there is debate: the coefficient is likely a function of grain size. We rule out solution 3 as a viable universal predictor because grain size is a control on ripple length (e.g., Cummings et al., 2009). We focus our remaining examination on the solution at complexity 8.

$$\lambda = \frac{d_0}{1.12 + 2.18(1000D_{50})}$$

Equation 2.9

Figure 9 shows the general behavior of this predictor: increasing wave ripple spacing with increasing bottom orbital excursion and decreasing wave ripple wavelength with increasing grain size.

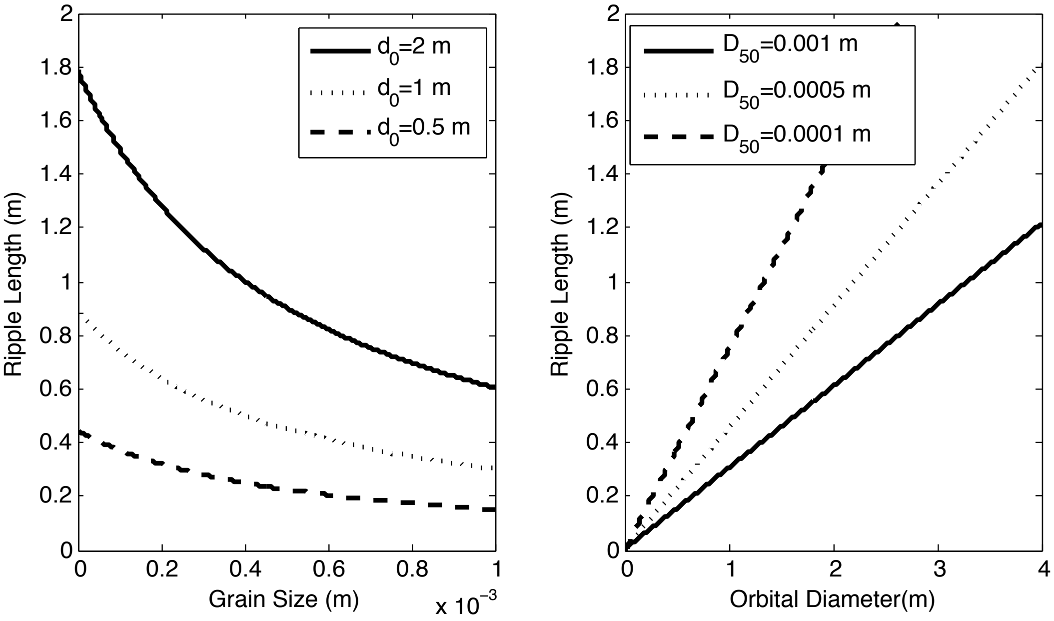


Figure 9: Example behavior of ripple length predictor as a function of grain size for given bottom orbital excursions (left panel) and as a function of bottom

orbital excursion for given grain size (right panel).

Furthermore ripple length is more sensitive to median grain size at larger orbital diameter. Previous ripple length prediction schemes have focused on orbital diameter and grain size as they represent the two fundamental length scales in the development of oscillatory bedform. For instance, Soulsby and Whitehouse (2005) develop an equilibrium predictor where A/ D_{50} is the controlling parameter after examining the collapse of compiled data with several other variables.

Using only the reserved testing data, the NRMSE of the new GP predictor as well as those developed by Soulsby and Whitehouse (2005) and Pedocchi and García (2009a), are 0.74, 1.33, and 1.22 respectively, and the correlation coefficient is 0.78, 0.02, and 0.20 respectively. The GP derived predictor performs better than the other predictors based on the NRMSE and correlation coefficient. Figure 10 shows the performance of these models in both linear and log-log space.

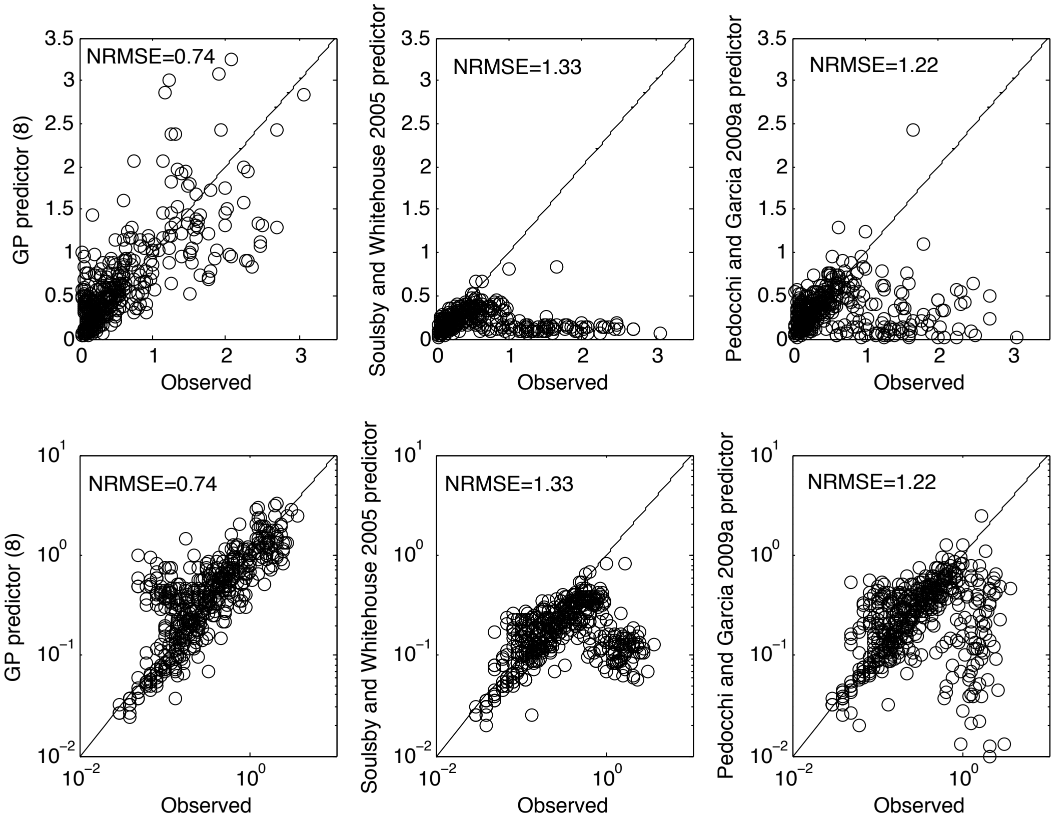


Figure 10: GP predictor of ripple length (2.8), Soulsby and Whitehouse (2005) predictor (2.3) and Pedocchi and García (2009a) predictor (2.5) evaluated using only the independent testing dataset. Top row shows the predictors in linear space, while bottom row shows log-log space.

Neither of these previously published predictors were developed for large scale orbital ripples, and both show predictions that deviate significantly when observed ripple wavelengths are large. The GP derived predictor is better able to capture large scale ripples. Both Soulsby and Whitehouse (2005) and Pedocchi and García (2009a) are able to better capture small scale 'anorbital' ripples that deviate significantly from the scaling of (2.9).

2.4.2 Ripple Height

The GP algorithm output is shown in Table 3.

Table 3: Solutions for Ripple Height

Solution	C	MSE
$\eta = 0.435 d_0$	3	0.0017
$\eta = 0.313\lambda(1000D_{50})$	5	0.0013
$\eta = \dfrac{\lambda(1000D_{50})^2}{0.372 + 5.29(1000D_{50})^2}$	14	0.0012
$\eta = \dfrac{\lambda(1000D_{50})^3}{0.0731 + 5.57(1000D_{50})^3}$	18	0.0012
$\eta = \dfrac{0.0237\lambda(1000D_{50}) + \lambda(1000D_{50})^3 - 0.308\lambda(1000D_{50})^2}{0.0332 + 4.46(1000D_{50})^3 - 0.321D_{50}}$	36	0.0010

This experiment evaluated 10^{10} formulas to develop the Pareto front shown in Figure 11.

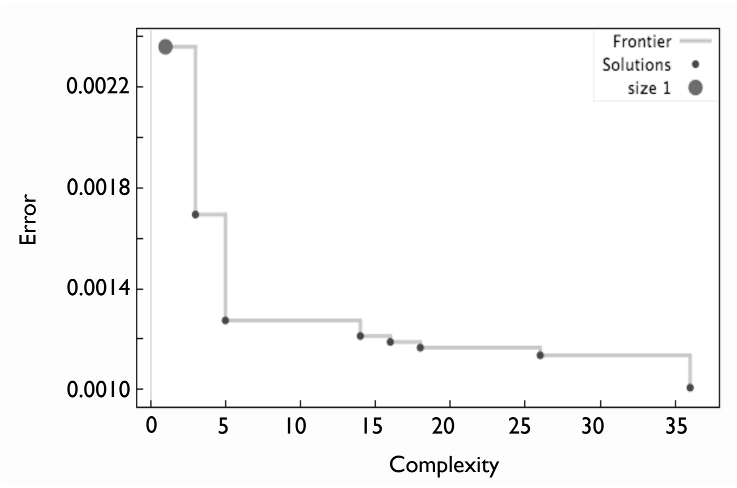


Figure 11: Pareto front for ripple height; Error is mean squared error of candidate solution versus the validation data set. Complexity is a quantification of the candidate solution length (both mathematical operators and variables).

Cliffs occur along the Pareto front at complexities of 3, 5, 14, 18 and 36 (Figure 11).

Predictor of complexity 3, $\eta = 0.435d_0$, is qualitatively similar to predictions of ripple height in the orbital regime (i.e. weak hydrodynamics) presented in Wiberg and Harris (1994), where ripple wavelength is a function of orbital diameter and ripple steepness (η/λ) is constant, therefore ripple height is a linear function of bottom orbital diameter. Constant steepness breaks down in stronger hydrodynamic conditions, and this is reflected in the inclusion of grain size and ripple length in more complex predictors. We have no compelling evidence to use the most nonlinear but best fit solution (36), nor is there compelling evidence at this time that ripple height has a such a strongly nonlinear dependence on grain size (Solution 14 and 18). We focus our analysis on solution 5:

$$\eta = 0.313\lambda(1000D_{50})$$

Equation 2.10

or, replacing λ (which denotes predicted ripple wavelength) with equation 9:

$$\eta = \frac{0.313d_0(1000D_{50})}{1.12 + 2.18(1000D_{50})}$$

Equation 2.11

Figure 12 shows the behavior of this predictor under conditions of various orbital diameter and grain size.

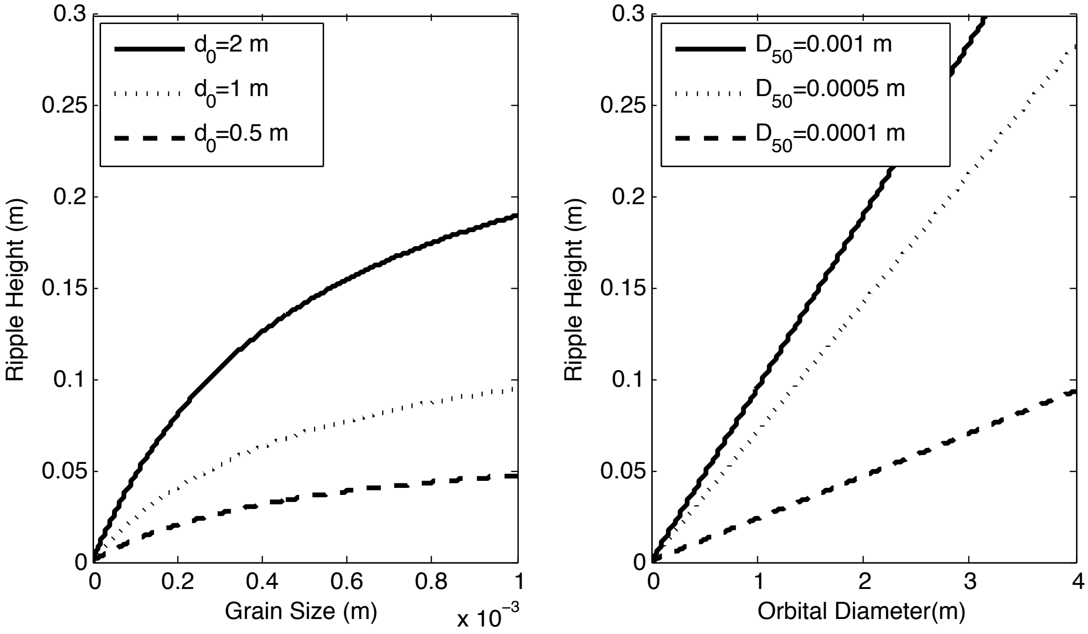


Figure 12: Example behavior of Ripple height predictor as a function of grain size for given bottom orbital excursions (left panel) and as a function of bottom orbital excursion for given grain size (right panel).

Ripple height increases with increasing grain size and orbital diameter. As with ripple length, ripple height is more sensitive to changes in grain size than changes in orbital velocity. Reserved testing data is used as an independent dataset to compare the GP predictor as well as those developed by Soulsby and Whitehouse (2005) and Pedocchi and García (2009a): the NRMSE for each predictor is 0.79, 1.02, and 1.01 respectively, and the correlation coefficient is 0.67, 0.41, and 0.47 respectively. The GP derived predictor performs better than the other predictors based on the NRMSE and correlation coefficient. Figure 13 shows the performance of these models in both linear and log-log space.

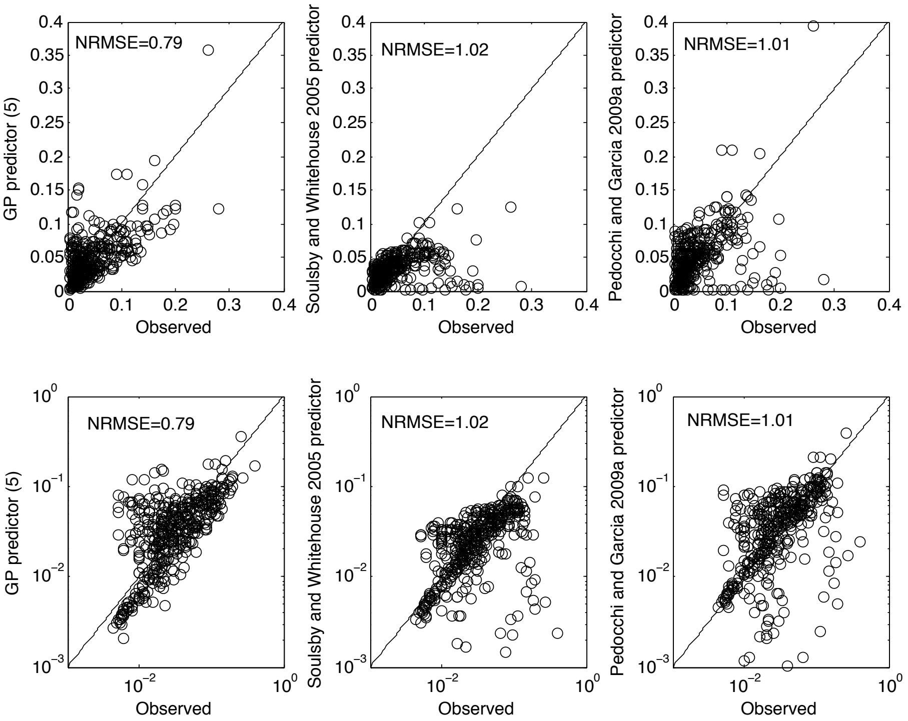


Figure 13: GP predictor of ripple height (2.10), Soulsby and Whitehouse (2005) predictor (2.3) and (2.4) and Pedocchi and García (2009a) predictor (2.6) evaluated using only the independent testing dataset. Top row shows the predictors in linear space, while bottom row shows log-log space.

2.4.3 Ripple Steepness

Combining the GP predictors for ripple length (2.9) and height (2.11), or simply rearranging (2.10), yields a predictor for ripple steepness:

$$\vartheta = 0.313(1000D_{50})$$

Equation 2.12

implying that steepness is a function solely of grain size, which is a gross approximation of the variability observed in the data, and to some extent even unphysical. To enhance our steepness prediction we produce a GP derived steepness predictor. The GP algorithm output is shown in Table 4.

Table 4: Solutions for Ripple Steepness

Solution	C	MSE
$\vartheta = 0.119$	1	0.0035
$\vartheta = 0.154 - 0.0613\lambda$	5	0.0026
$\vartheta = \frac{(1000D_{50})}{\lambda + 6.23(1000D_{50})}$	8	0.0021
$\vartheta = \frac{3.42}{22 + \left(\frac{\lambda}{(1000D_{50})}\right)^2}$	10	0.0019
$\vartheta = \frac{0.447}{2.81 + \left(\lambda^2 + \frac{-0.617\lambda}{(1000D_{50})}\right)^2}$	16	0.0017

This experiment evaluated 10^{10} formulas to develop the Pareto front shown in Figure 14.

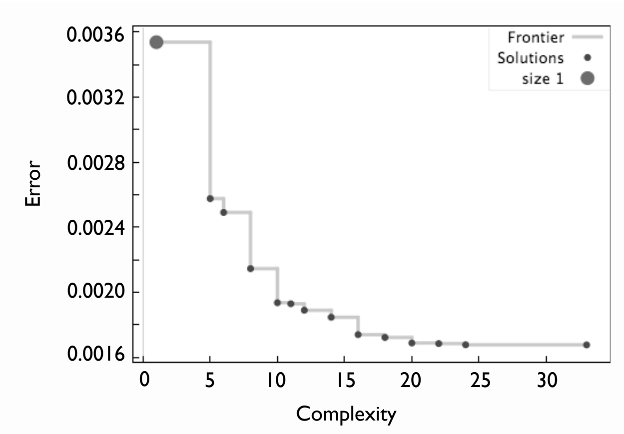


Figure 14: Pareto front for ripple steepness; Error is mean squared error of candidate solution versus the validation data set. Complexity is a quantification of the candidate solution length (both mathematical operators and variables).

Cliffs, significant gains in error for small changes in complexity occur along the Pareto front at complexities 5, 8, 10, and 16. The predictor at complexity 5 produces nonphysical results (negative steepness under some conditions) so is ruled out. The most nonlinear predictor reported (complexity of 16) shows only small decrease in error for increasing equation complexity; we focus our analysis on predictor 10:

$$\vartheta = \frac{3.42}{22 + \left(\frac{\lambda}{(1000D_{50})}\right)^2}$$

Equation 2.13

by replacing λ (predicted ripple wavelength) with (2.9), yields:

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

Equation 2.14

Figure 15 shows the behavior of this predictor under conditions of various orbital diameter and grain size.

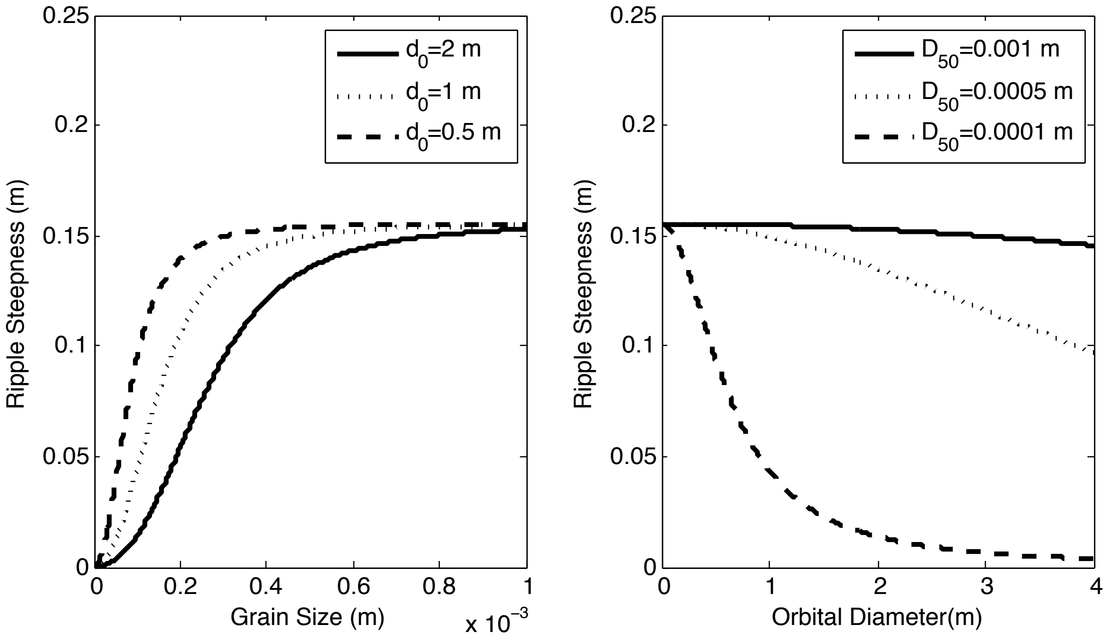


Figure 15: Example behavior of Ripple Steepness predictor as a function

of grain size for given bottom orbital excursions (left panel) and as a function of bottom orbital excursion for given grain size (right panel).

Increasing D_{50} (for a given d_0) results in increasing ϑ until a saturated value of 0.15 is reached. Increasing d_0 (for a given D_{50}) results in decreasing ϑ. Small grain sizes are very sensitive to changes in d_0, while large grain sizes are relatively insensitive. Figure 16 shows the performance of (2.14) against the independent testing data compared to the linear convolution of GP derived length and height (2.12), as well as the Pedocchi and García (2009a) and Soulsby and Whitehouse (2005) predictors.

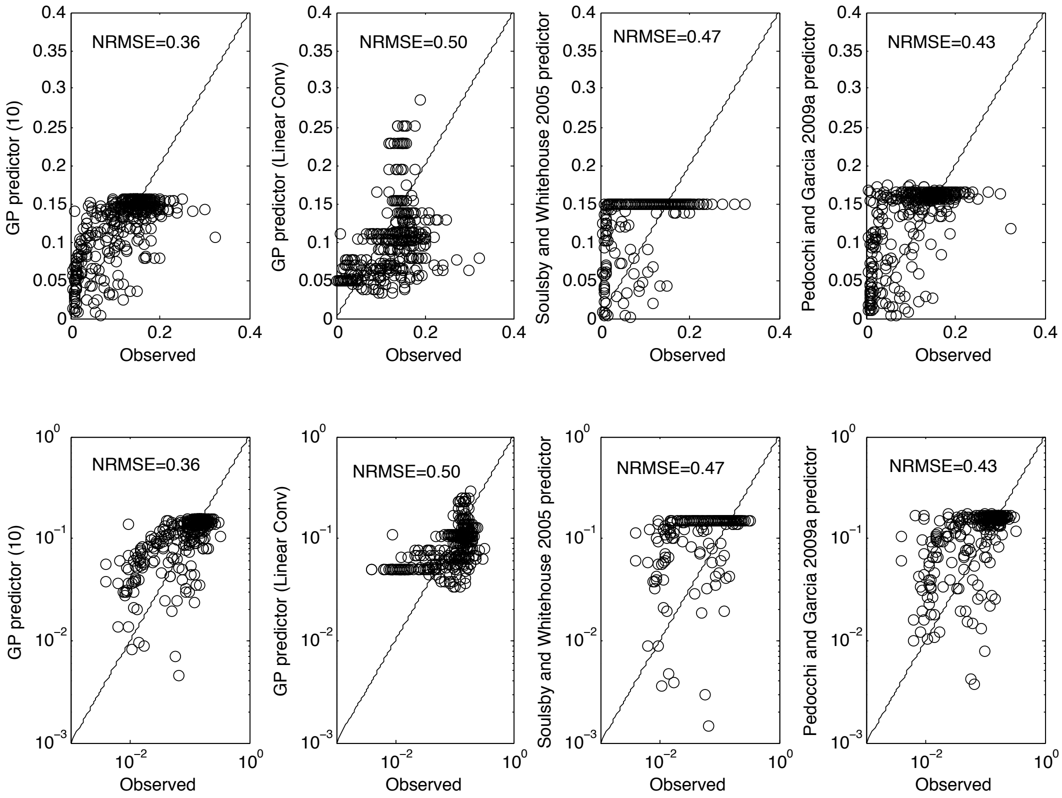


Figure 16: GP predictor of ripple steepness (2.13), Predictor based on linear convolution of GP height and length (2.11), Soulsby and Whitehouse (2005) predictor (4) and Pedocchi and García (2009a) predictor (2.5) and (2.6) evaluated using only the independent testing dataset. Top row shows the predictors in linear space, while

bottom row shows log-log space.

The NRMSE of these predictors is: 0.36, 0.50, 0.47, and 0.43, respectively, and the correlation coefficient is 0.70, 0.48, 0.63, and 0.50 respectively. The GP derived predictor performs better than the other predictors (including the linear convolution of GP derived λ and η) based on the NRMSE and correlation coefficient.

2.5 Discussion

2.5.1 Predictors derived from Genetic Programming

The suite of predictors that are produced as output of the genetic programming show a trend of increasing predictability with increasing complexity. Highly nonlinear predictors have been avoided in this study because they may be fit to the noise or variance present in the training dataset (i.e. they are overfit). Yet the more complex nonlinear predictors can be used as hypothesis for further field and lab studies where grain size effects are a focus.

Dependence on orbital scaling and grain size is not imposed by the authors, it is a result of the data used to feed the genetic programming software. Aside from the data sets used in this study, other field observations have shown decreasing ripple height and increasing ripple length in fine grained sand under strong hydrodynamic forcing (e.g., a transition from steep to low profile bedforms; Green and Black, 1999; Green et al., 2004; Trembanis et al., 2004). Pedocchi and García (2009a) and Cummings et al. (2009) note that U is the major control on the transition from small ripples (anorbital) to large ripples (hummocks), yet our GP derived predictors contain only one hydrodynamic parameter, d_0. Furthermore dependence on U is not present in any of the candidate predictors (Tables 2, 3, and 4). This is likely the result of several factors: First, d_0 and U are correlated in our database (Figure 5), making d_0 a potential proxy for any dependence on U. Second, our database likely contains multiple ripple sizes at similar hydrodynamic conditions, resulting in the lack of a clear velocity threshold. Third, we focus on developing a continuous predictor so do not include any logical statements that can accommodate a threshold.

Our results show that ripple height is more difficult to predict than ripple length (e.g., Yan et al 2008; Williams et al 2004). As mentioned previously, this is likely a consequence of ripple crests being subject to higher flow velocities and suspension processes. Yet successful height and steepness determination is important for the prediction of sediment transport, in particular the reference concentration (e.g., Green and Black 1999) and sediment diffusivity (e.g., Nielsen 1992, Thorne et al., 2009). Only 2 equations are needed to predict height, length and steepness of ripples, but error in the two chosen predicted parameters cascades to the third. The basic linear convolution of predicted λ (2.9) and predicted η (2.11) demonstrate this cascading error: the resultant steepness predictor (2.12) produces results that are solely dependent on grain size. We instead offer 3 separate equations in the hope that workers will decide which 2 predictors are most valuable for a specific research question. Notably, the GP algorithm did have predicted λ and predicted η available as equation building blocks when determining ripple steepness but the term 'η/λ' did not appear in any candidate solutions (Table 4). Generating 3 separate predictors that are not self-consistent leads to geometric inconsistencies, but results in better prediction for work that requires accurate prediction of height and steepness but does not rely on ripple length measurements.

The hydrodynamic and sedimentological limit of the current prediction scheme is represented by the 4-dimensional shape that outlines the point cloud in Figure 5. (Table 1 contains more information regarding the range of the dataset). We excluded conditions where ripples are not present either as a result of sheet flow conditions (upper plane bed) or because of insufficient mobility (lower plane bed). Uncertainty in the onset of upper plane bed exists because of the lack of data at a range of D_{50} in field-scale conditions (e.g., Li and Amos, 1999; Trembanis et al., 2004; You and Yin, 2006). Additionally, field work suggests that upper plane bed conditions may not be flat, but instead typified by dynamic features that may be similar if not identical to long wave ripples (Green and Black, 1999; Green et al., 2004; Trembanis et al., 2004). As a result of the ambiguity in bed state under 'upper plane bed' conditions we did not compare our predictor to the version developed by Camenen (2009), which explicitly includes a sheet flow threshold (where ripples are destroyed). Furthermore we do not compare the GP derived predictors with those developed by Williams et al., (2005), who developed separate predictors for short wavelength and long wavelength ripples. We intentionally did not divide our dataset (and develop separate predictors) in an attempt to construct a practical prediction scheme that spans a wide range of conditions. Since this study aims to produce a continuous predictor of wave ripple geometry, the use of discontinuous functions (logical statements: e.g. 'if-then-else') has not been explored quantitatively in this contribution. This study does not tackle the issue of time dependent adjustment of bedforms in unsteady flow (Austin et al., 2007, Soulsby et al., 2012, Traykovski 2007; Davis et al., 2004; Doucette and O'Donoghue, 2006; Hay 2008), the importance of initial conditions on final ripple configuration (Traykovski et al., 1999; Hansen et al., 2001), or the explicit incorporation of emergent ripple parameters (e.g., defect density; Skarke and Trembanis, 2011).

2.5.2 Open Research Questions: Wave Ripples

Data integration campaigns can highlight gaps in knowledge. The collection of new ripple datasets will be able to be used as either independent tests of the predictors developed in this study (if the setting corresponds to an area in figure 5 that is dense with points) or as new data to train the GP algorithm (if the data correspond to unexplored or sparse area in figure 5; Bowden et al., 2012). Additional datasets of wave ripple geometry that include more input parameters (e.g. measures of grain sorting, wave irregularity, initial conditions, time dependence) are needed if prediction accuracy is to increase. Furthermore datasets that encompass coarse grained environments (coarse sand and gravel) and datasets in energetic conditions are still needed. Though coarse grained conditions reflect a smaller fraction of the seabed than fine grained settings, coarse grained environments are likely important nursery habitat for fish (Hallenbeck et al., 2012). Collection of this data will not only help in the determination of ripple configuration under these specific forcing conditions, but linking these environments with the present data will allow for the development of a better predictor by defining the shape of the prediction surface over a greater extent in phase space.

Conditions with waves and currents (e.g., Lacy et al., 2007; Khelifa and Ouellet, 2000; Arnott and Southard, 1990) are excluded from this analysis. The collection and integration of data with waves and currents may lead to a more universal bedform predictor in the future but more data on ripple geometry under wave and current forcing is needed for machine learning techniques to be applied successfully. The dataset in this study uses median grain size as the sole sedimentological metric for predicting ripple geometry. Yet many field settings may not be accurately described by a sharp peaked unimodal distribution of grain sizes, and therefore prediction of ripples using D_{50} may lead to significant error. Foti and Blondeaux (1995) showed that the addition of coarse sediment can act as stabilizing feature, enhancing ripple length. It is possible that graded sediment will not conform to the predictive tools outlined above. But what is the effective D_{50} in graded sediment when predicting ripples? Furthermore variations in grain shape and bed porosity may also impact the geometry of ripples. More research is needed into the role of mixed grains in determining equilibrium wave ripple geometry (e.g. Calantoni et al., 2013).

More studies are needed to better constrain thresholds between short ‘anorbital’ ripples and large ‘orbital’ scale ripples (Pedocchi and García, 2009; Cummings et al., 2009; Maier and Hay, 2009). Experimental work has thus far shown that there exists no intermediate scale between these two configurations (Dumas et al., 2005; Cummings et al., 2009). The determination of when large ripples appear and when superimposed short ripples disappear will allow the pruning of the database in regions where overlapping ripple scales occur. The decision of which ripple scale to eliminate when both exist is a function of the research question being studied.

2.5.3 Open Research Questions: Data Driven Prediction

In this contribution we demonstrate a selection technique whereby very few data are used to train the GP algorithm and most data is used as validation and independent testing. The training data was selected solely from variables representing the forcing conditions. As a result the training data is not representative of the entire population of ripple configurations as data points that are neighbors in 'forcing space' do not necessarily have similar ripple geometries. The selected training data is therefore only related to the range of forcing present in the dataset, not the range of ripple geometry. Therefore we believe that our sampling strategy does not bias the testing of the predictors (which relies on ripple geometry) using the reserved, unselected testing data.

We define the testing dataset as ‘independent’ because it was not shown to the

GP algorithm. Additionally we performed experiments by removing several individual datasets from the composite dataset. The removed data serves as testing data that is not shown to the selection routine, not shown to the GP algorithm, and additionally not related to data shown to the selection routine/GP (this is another definition of 'independent'). The resultant predictors (not shown) were quantitatively similar to those presented in this contribution and similarly performed better than the Pedocchi and García (2009a) and Soulsby and Whitehouse (2005) predictors using only the smaller sample of removed datasets as testing. Is it enough for testing data to be unseen by the ML algorithm, or do entire datasets need to be reserved whole as testing data? More investigation will resolve this issue.

Even though we were able to obtain good results using few centroids, we are unaware of a technique for quantitatively determining the optimal number of centroids to capture the variability in the data set while leaving the maximum amount of data for use as validation/testing. Furthermore many selection and clustering routines are available, and it is unclear which routine is optimal for a given dataset. It is likely that some of the answers to these questions lie in statistical science and computer science literature that has not fully percolated into the Earth Sciences.

Our observations with the GP software show too few centroids tend to underfit the data because the GP has too little training data to develop applicable solutions. With few training data solutions tend to be linear and have low RMSE when compared tot the validation and testing datasets. Training datasets that are larger than used in this study (> 40 centroids) tend to produce more large (complexity > 30) nonlinear solutions. In addition, the solutions at complexity less than 30 are similar (if not identical) to the solutions in this study (using smaller training datasets). The invariance of solutions gives qualitative justification to the number of centroids used in this study, but we do not offer a quantitative technique for determining the minimum number of training data needed to capture dataset variability. Furthermore how is this number linked to the quantity and quality of the data in the training dataset? Lastly, the stability and final criterion for selecting a single predictor is subjective and can likely be improved or quantitatively justified by implementing more sophisticated accounting techniques based on information such as the Aikake Information Criterion (AIC) or the Bayesian Information Criterion (BIC).

2.6 Conclusion

We develop equilibrium predictors of oscillatory ripple geometry using genetic programming. Ripple length is a weak nonlinear function grain size and bottom orbital excursion. Ripple height and steepness are nonlinear functions of grain size and predicted ripple length (i.e. grain size and bottom orbital excursion). Furthermore these new predictor encompass a wide range of hydrodynamic and sedimentological conditions not previously included in published prediction schemes. However, the proposed method is not suitable for practical applications with significant currents present, nor under conditions that would either be below the threshold of motion or above the threshold of ripple wash-out. Such conditions should be identified separately by existing methods (Nielsen, 1992; Lacy et al., 2007; Camenen, 2009; Soulsby et al., 2012)

This contribution further demonstrates the viability of developing empirical predictors through ML techniques. As previously mentioned by Oehler et al., (2012), ML algorithms could be integrated into future morphodynamics models (model-data fusion and the development of a 'hybrid' model; Krasnapolsky and Fox-Rabinovitz, 2006), replacing functions with large uncertainty.

The data integration campaign (which preceded the implementation of the GP algorithm) had the side benefit of highlighting the current state of our knowledge on ripple geometry, potentially motivating targeted data collection campaigns. Newly collected data can be fed back into the GP software to develop revised predictors.

3. Data Driven Components in a Model of Inner Shelf Sorted Bedforms: A New 'Hybrid' Model

Goldstein, E. B., G. Coco, A. B. Murray, and M.O. Green, 2014. Data-driven components in a model of inner-shelf sorted bedforms: a new hybrid model, Earth Surface Dynamics, 2, 67–82.

3.1 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 hydro- dynamic 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 dataset) 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 datasets 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 datasets with nonlinear relationships and multiple independent variables are difficult to visualize and interpret. Traditional techniques for developing successful parameterizations include converting multidimensional datasets 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 datasets 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.

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 (km scale) patches of segregated coarse and fine grained sediment (Figure 17) with only slight bathymetric relief (cm–m scale) relative to bedform pattern wavelength (10 m–km) are present on many continental shelf systems (Coco et al., 2007b).

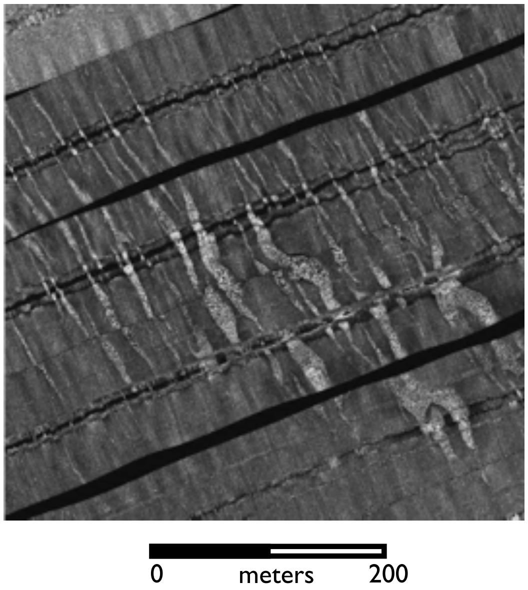


Figure 17: Sorted bedforms present in ~ 5 m of water off the coast of Tairua Beach, New Zealand (Coco et al., 2007a). White areas are composed of coarse sediment, while dark areas are floored by fine sediment. Shoreline is towards the bottom of the

panel.

Unlike most bed forms 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 “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 sea bed 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. Coco et al. (2007b) did show the potential for coarse domains to occur in the trough of bedforms, but 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 dataset of Green and coworkers (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.

3.2 GP Methods

3.2.1 Dataset

Figure 18 shows the multi-setting field dataset composed of 1748 individual measurements from 6 separate field experiments at different locations in New Zealand.

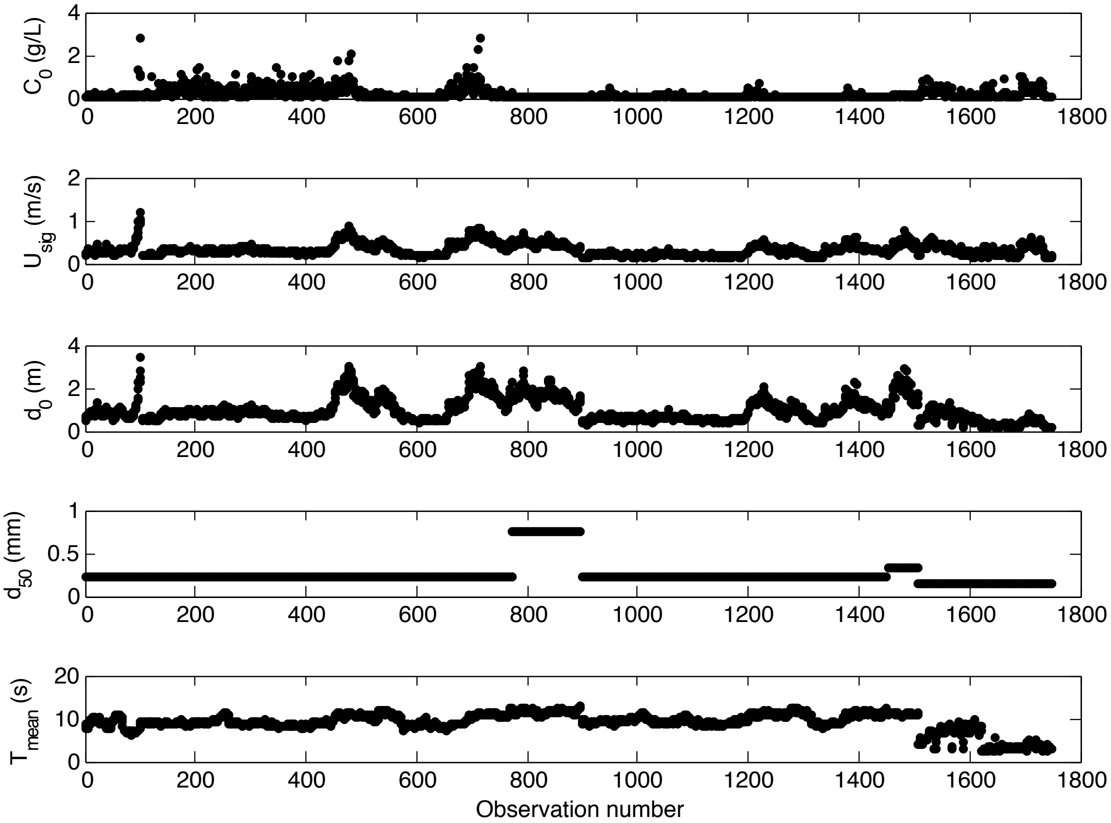


Figure 18: Observations of suspended sediment reference concentration data set C_0 and concomitant measurements of significant wave velocity at the bed (U_{sig}), wave orbital excursion at the bed (d_0), mean grain size of bed material (d_{50}), and mean spectral wave period at the bed (T_{mean}). Note that mean grain size of bed material is shown here in millimeters. A similar figure appears in Oehler et al. (2012).

We briefly summarize the experiments below and in Table 5; 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.

Table 5: 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

A single experiment (Green and Black, 1999; Green, 1999) collected 127 measurements seaward of the surfzone with mean water 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 dataset is ideal for application in the sorted bedform model because three of the six experiments in the composite dataset are derived from a sorted bedform field (Green et al., 2004; Trembanis et al., 2004).

3.2.2 Selection of Training, Validation and Testing datasets

The database is split into three subsets to be used as training, validation, and testing. The training dataset is used to develop candidate solutions. The validation dataset 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 dataset 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. 19), the selection and partitioning of data into these three categories is crucial to develop a well performing predictor applicable to a range of environments (e.g., Bowden et al., 2002).

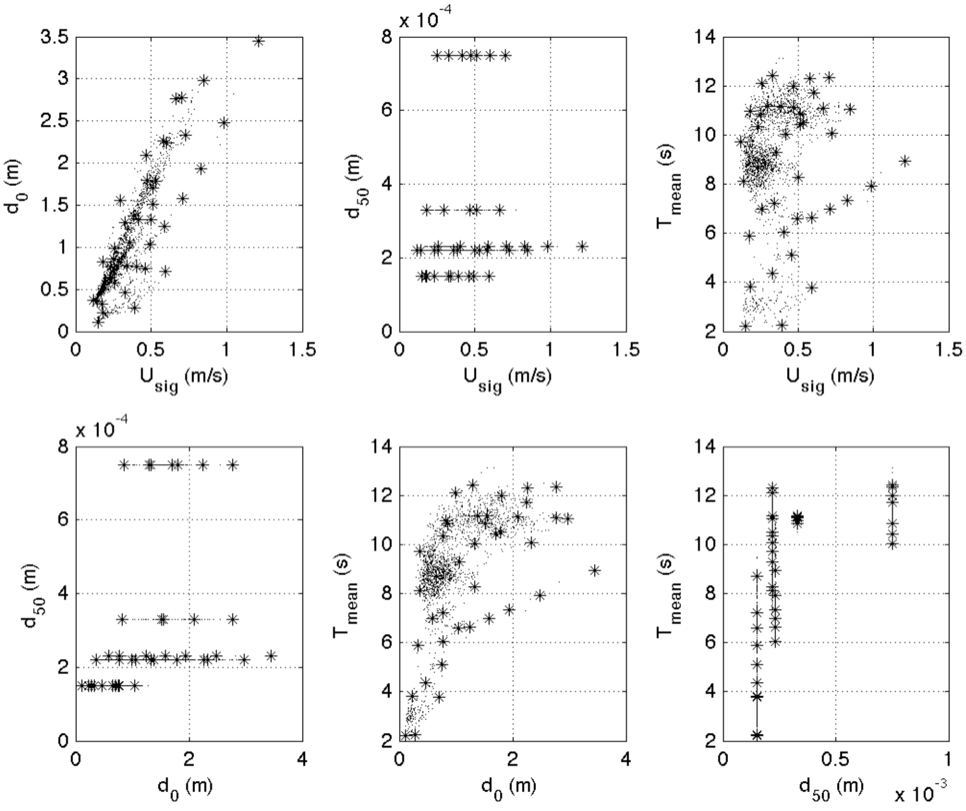


Figure 19: Visualization of the range of conditions in the C_0 data set. Each plot represents a two-dimensional projection of the entire data set onto the set of axes shown. For instance, the first panel with data projected onto the d_0 – U_{sig} plane shows no information about d_{50} or T_{mean}. Stars denote centroid locations (training data), while points denote unselected data (validation and testing). Note that centroids are distributed throughout the data set.

The C_0 dataset 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 is randomly divided, there is a potential that the training data excludes data from sparse regions in the dataset (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 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 dataset 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 4-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 is then denormalized.

There remains significant ambiguity in determining the appropriate number of centroids needed to accurately represent a dataset, especially continuous data (e.g., May et al., 2010; Goldstein et al., 2013). Selecting too many centroids can rob the validation and testing datasets 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., Gonlçaves 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 dataset (Goldstein et al., 2013). We use 40 centroids for the prediction of C_0 (centroid locations can be seen in Fig. 18.), 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 dataset is split between validation and testing randomly, without using a selection routine. The final breakdown for the datasets is ~ 2 % training, ~ 49 % validation, ~ 49 % testing.

3.2.3 Genetic Programming

We operate on this dataset 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, 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), √ (square root), as well as integer powers ($e.g.\ \ x^2, x^3, x^4$). 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:

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

Equation 3.1

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. 20).

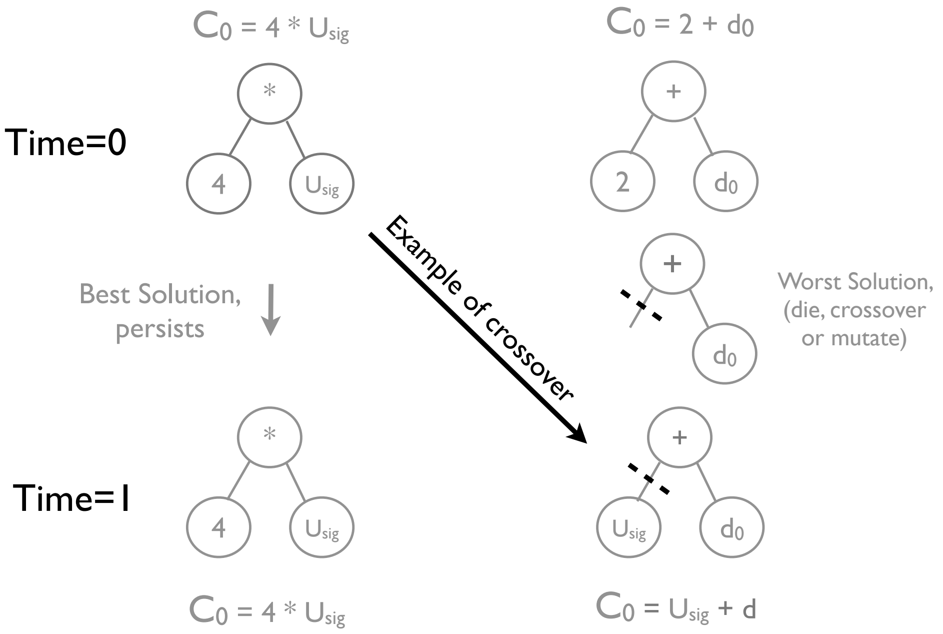


Figure 20: Example of the genetic programming process. Potential solutions are encoded as a population of trees. Here a hypothetical population of two solutions is shown. The first solution has a low MSE and therefore persists to the next iteration. The second solution has a high MSE and therefore is 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.

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 overfit, predictors). Several methods exist for eliminating overfit solutions (e.g., Gonçalves et al., 2012). We use several 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 over- fit (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. 21).

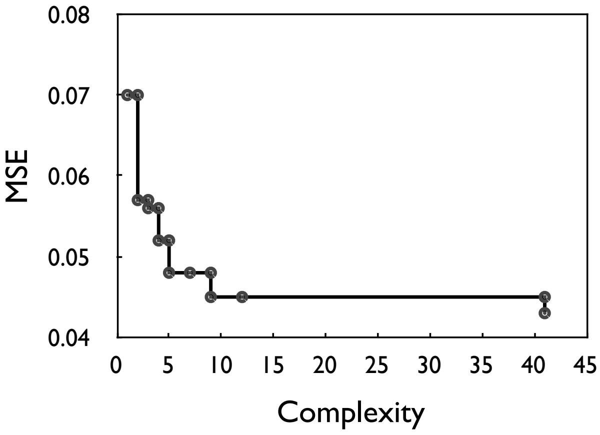


Figure 21: Reference concentration Pareto front; MSE is mean squared error of

candidate solution versus the validation data set. Complexity is a quantification of the candidate solution length (both mathematical operators and variables).

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.

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):

$$\mathrm{NRMSE} = \frac{\sqrt{\mathrm{MSE}}}{\bar{\mathrm{b}}}$$

Equation 3.2

where b is the mean of the observed values. Additionally we report 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.

3.3 GP Results

The GP algorithm output is shown in Table 6 (Note that numerical coefficients listed in the table are dimensional).

Table 6: Solutions for reference concentration

Solution	Complexity	MSE
$C_0 = 0.182$	1	0.070
$C_0 = U_{sig}^2$	2	0.057
$C_0 = 0.637U_{sig}$	3	0.056
$C_0 = \left(1.19U_{sig}\right)^2$	4	0.052
$C_0 = U_{sig} - 0.647(1000d_{50})$	5	0.048
$C_0 = \left(\frac{0.235U_{sig}}{(1000d_{50})}\right)^2$	7	0.048
$C_0 = \left(\frac{0.328U_{sig}}{0.0688 + (1000d_{50})}\right)^2$	9	0.045
$C_0 = \left(1.27\sqrt{U_{sig}} - 1.21(1000d_{50})\right)^2$	12	0.045
$C_0 = \frac{0.179U_{sig}^2 - 0.00538}{d_0(1000d_{50})} + \frac{0.185 + 0.179U_{sig}^2 d_0 - 0.179U_{sig}^2 - 0.0319U_{sig}^4}{(1000d_{50})}$	41	0.043

This experiment evaluated 10^{10} formulas to develop the Pareto front shown in Fig. 21. Cliffs occur along the Pareto front at complexities of 2, 4, 5, 9, and 41 (Fig. 21). 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):

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

Equation 3.3

Note that the coefficients of Eq. (3.3) are dimensional. Reserved testing data is used as an independent dataset to 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. 22.

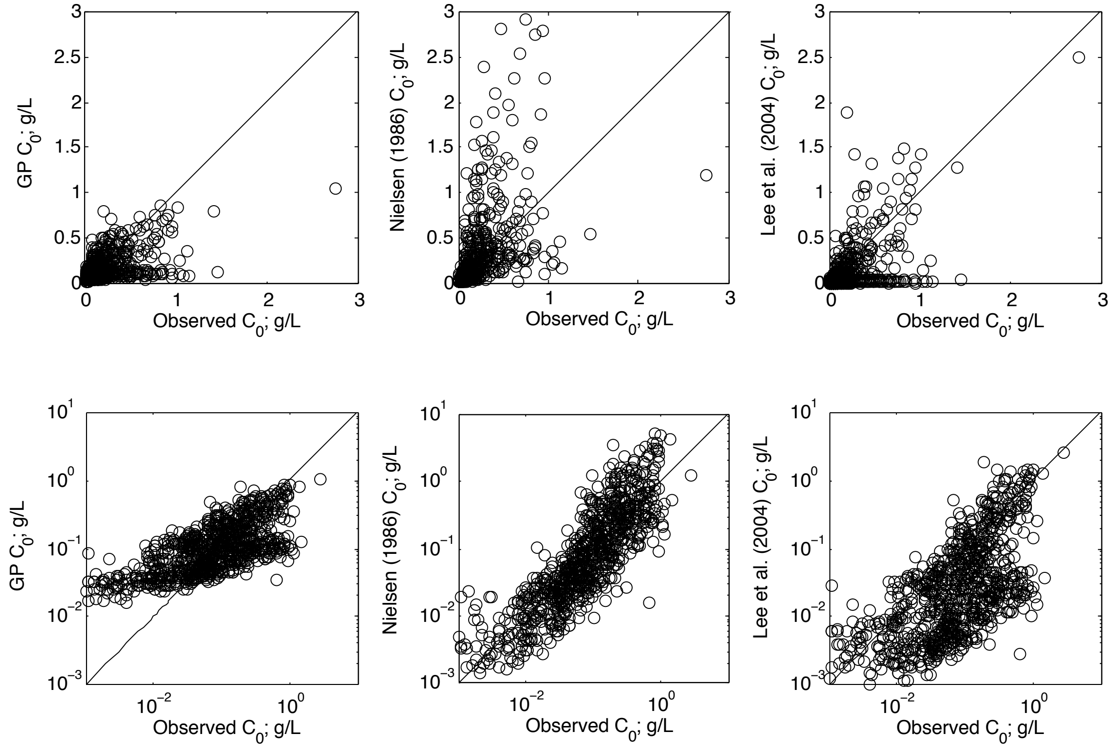


Figure 22: GP predictor of *C*₀, Nielsen (1986) and Lee et al. (2004) predictor evaluated using only the independent testing data set. Top row shows the predictors in linear space; bottom row shows log–log space.

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. (3.3) deteriorates.

3.4 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 trans- port 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_{coarse} = 0.0005 m and d_{fine} = 0.0002 m; fall velocity w_{coarse} = 0.07 m s^{-1} and w_{fine} = 0.02 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 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 pro- files (Murray and Thieler, 2004; Coco et al., 2007a). The flux of suspended sediment ($q_{susp,s}$),

$$\vec{q}_{susp,s} = \int C_s \vec{V} dz - \gamma_s \frac{1}{5W_s} U_w^5 \vec{\nabla} z$$

Equation 3.4

$$\gamma_s = \gamma_c \frac{16E\rho}{3\pi w_s} C_d$$

Equation 3.5

where U_w is the maximum wave orbital speed at the bed (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. (3.4) 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. (3.4), similar to the methodology of Calvete et al. (2001). For all experiments in this contribution $\gamma_c = 0.07$. 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:

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

Equation 3.6

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):

$$z_0 = \frac{1}{30}(2d_{50} + 28\eta\vartheta)$$

Equation 3.7

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):

$$C_s(z) = C_{0,s} e^{-\frac{w_s z}{\varepsilon_s}}$$

Equation 3.8

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 $U_{sig} = U_w$ and $d_{50} = d_s$ and therefore Eq. (3.3) becomes:

$$C_0 = \left(\frac{0.328U_w}{0.0688 + (1000d_s)}\right)^2$$

Equation 3.9

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):

$$\varepsilon_s = \Omega k_s U_w$$

Equation 3.10

$$k_s = 25\eta\vartheta$$

Equation 3.11

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):

$$\eta = \frac{0.313d_0(1000D_{50})}{1.12 + 2.18(1000D_{50})}$$

Equation 3.12

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

Equation 3.13

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

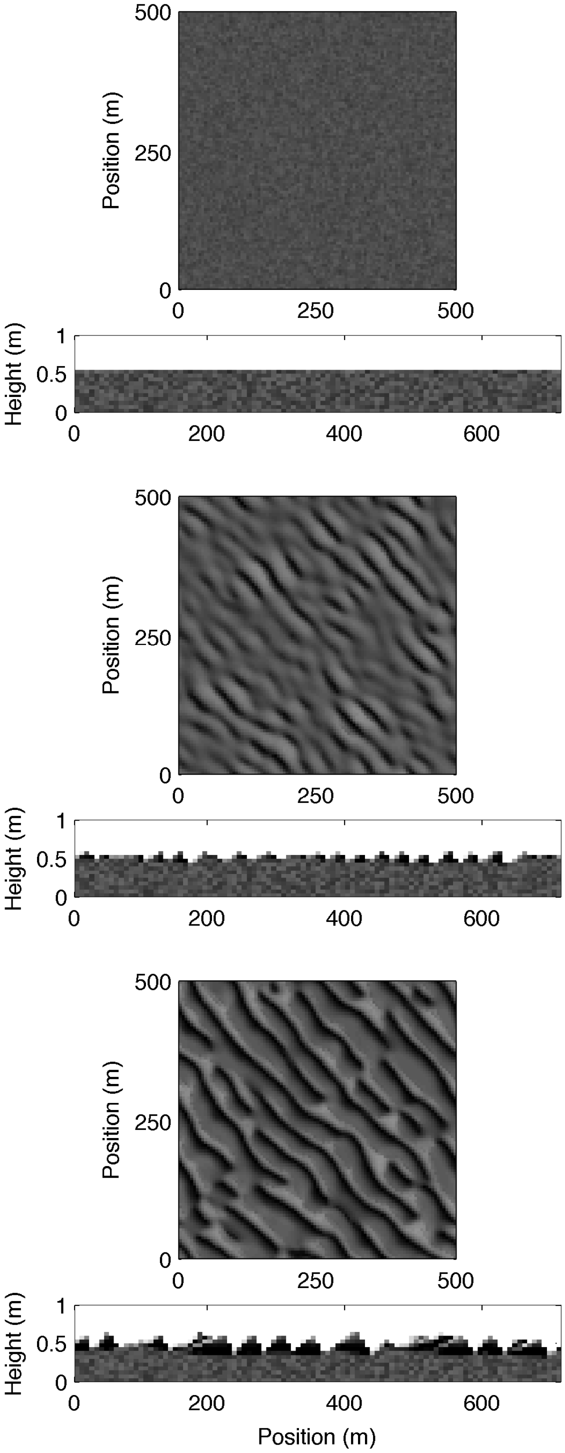
$$d_{50,i} = \left(1 - B_{coarse,i}\right)d_{fine} + B_{coarse,i}d_{coarse}$$

Equation 3.14

where $B_{coarse,i}$ 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.

3.5 Hybrid Sorted Bedform Model Results

The initially flat well mixed conditions can be seen in Fig. 23.



Position (m)
500
250
0
0
250
500
Height (m)
1
0.5
0
0
200
400
600
Position (m)
500
250
0
0
250
500
Height (m)
1
0.5
0
0
200
400
600
Position (m)
500
250
0
0
250
500
Height (m)
1
0.5
0
0
200
400
600
Position (m)

Figure 23: Plan view and profile view of sorted bedform model output (note the vertical exaggeration of profile view). Black and white pixels indicate fine (d_{fine}= 0.0002 m) and coarse (d_{coarse} = 0.0005 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 bed- forms; note that coarse domains appear on the updrift flank of the bedforms and wavelength and height are relatively small

This configuration is unstable, and sorted bedforms emerge within 50 model days to form the rhythmic segregated pattern shown in Fig. 23. 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. Bed- forms 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 inter- actions: 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. 23. 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 are smaller), 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_{fine} = 0.0002 m; w_{fine} = 0.02 m s^{-1}) while the coarse fraction diameter is varied between 0.0003–0.001 m (w_{coarse} = 0.04–0.12 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. 24 (sorted bedform wavelength and height are evaluated after 100 model days).

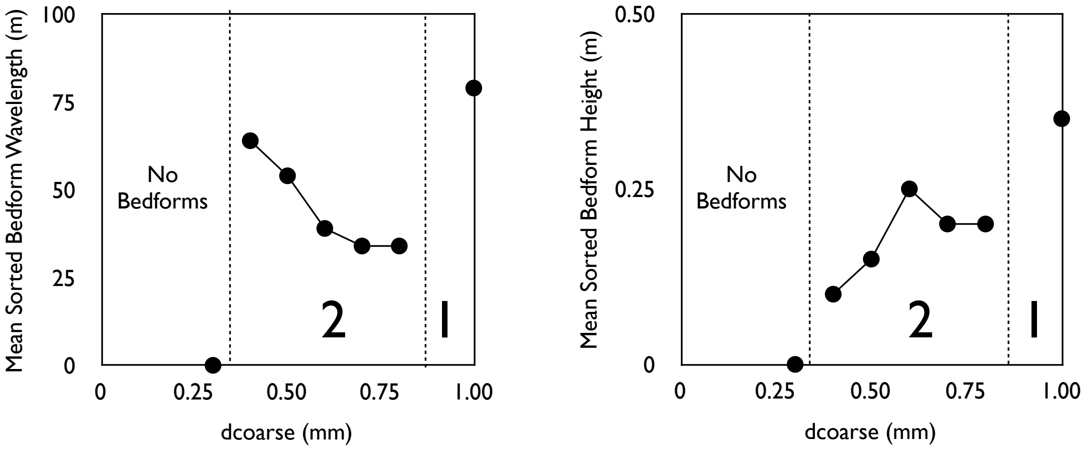


Figure 24: Variations in sorted bedform characteristics (wavelength and height) after 100 days when coarse grain size is held constant. No bedforms appear when the coarse material is too fine. Mode 1 bedforms (long wavelength, larger relief, coarse domains in trough) appear when coarse grain size is 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_{coarse} = 0.9 mm.

Similar to Coco et al. (2007a) sorted bedforms do not appear when the grain size contrast between size fractions is too small ($d_{fine}/d_{coarse} < 0.5$). When coarse grains range from 0.004–0.008 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 up- stream (Fig. 25).

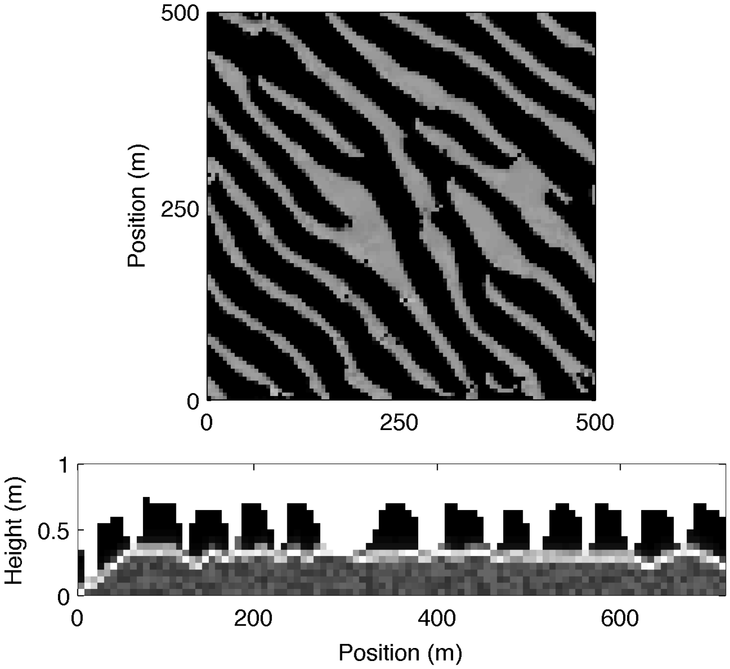


Figure 25: Plan view and profile view of mode 1 sorted bedforms after 50 days. Conditions are identical to Fig. 23 except d_{coarse}= 0.001 m. From identical initial conditions sorted bedforms appear much faster and are prominent features by 50 model days. Note that coarse domains appear solely in the bathymetric trough of the bed- forms and wavelength and height are relatively large.

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. (3.4). 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 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. 25). 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 wave- length 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.)

3.6 Discussion

3.6.1 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 datasets 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 $U_{sig} = U_w$ 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 $U_{sig} = U_w$ 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. (3.3) 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 nearbed 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 was 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. 22). 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 as 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).

3.6.2 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 over-predicts sediment transport; Fig. 22) 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 results 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 feed- back 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 up- drift 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 bathymetric-flow 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. 17). 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 success- fully calibrated. Calibration is accomplished by adjusting the variable γ_C in the morphodynamic diffusion term, Eqs. (3.4) and (3.5). 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 $\gamma_C = 0.05\text{-}0.08$. 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., $\propto U_w^5$) 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. (3.4) 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 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. (3.4) 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 if 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.

3.7 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)

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.: To- ward representing wave-induced sediment suspension over sand ripples in RANS models, J. Geophys. Res.-Oceans, 118, 1–15, 2013.

Andreotti, B., Claudin, P. and Pouliquen, O.: Aeolian Sand Ripples: Experimental study of Fully Developed States, Physical Review Letters v. 96, 2006.

Ardhuin, F., Drake, T.G., and Herbers, T.H.C.: Observations of wave-generated vortex ripples on the North Carolina continental shelf. Journal of Geophysical Research 107(C10), 3143, doi:10.1029/2001JC000986, 2002.

Arnott, R.W., and Southard, J.B.: Exploratory flow-duct experiments on combined-flow bed configurations, and some implications for interpreting storm-event stratification. Journal of Sedimentary Research 60(2), 211-219, 1990.

Austin, M.J., Masselink, G., O'Hare, T.J., and Russell, P.E.: Relaxation time effects of wave ripples on tidal beaches. Geophysical Research Letters 34, L16606, doi:10.1029/2007GL030696, 2007.

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. Journal of Hydraulic Research 45, 435-450, 2007.

Becker, J.M., Firing, Y.L., Aucan, J., Holman, R., Merrifield, M., and Pawlak, G.: Video-based observations of nearshore sand ripples and ripple migration. Journal of Geophysical Research 112, C01007, doi:10.1029/2005JC003451, 2007.

Blondeaux, P.: Mechanics of coastal forms. Annual Review of Fluid Mechanics 33, 339-370, 2001.

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.

Coastal Engineering 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 Resources Research 38(2), doi:10.1029/2001WR000266, 2002.

Bowden, G.J., Maier, H.R., and Dandy, G.C.: Real-time deployment of artificial neural network forecasting models: Understanding the range of applicability. Water Resources Research 48 W10549, doi:10.1029/2012WR011984, 2012.

Boyd, R., Forbes, D.L., and Heffler, D.E.: Time-sequence observations of wave-formed sand ripples on an ocean shoreface. Sedimentology 35, 449-464, 1988.

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

Calantoni, J., Landry, B.J. and Penko, A.M.: Laboratory observations of sand ripple evolution using bimodal grain size distributions under asymmetric oscillatory flows In: Conley, D.C., Masselink, G., Russell, P.E. and O'Hare, T.J. (eds.), Proceedings 12th International Coastal Symposium (Plymouth, England), Journal of Coastal Research, Special Issue No. 65, pp. 1497-1502, ISSN 0749-0208, 2013.

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.

Camenen, B.: Estimation of the wave-related ripple characteristics and induced bed shear stress. Estuarine, Coastal and Shelf Science 84, 553-564, 2009.

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

Chang, Y.S., and Hanes, D.M.: Suspended sediment and hydrodynamics above mildly sloped long wave ripples. Journal of Geophysical Research 109, C07022, doi:10.1029/2003JC001900, 2004.

Charru, F., and Hinch, E.J.: Ripple formation on a particle bed sheared by a viscous liquid. Part 2. Oscillating flow. Journal of Fluid Mechanics 550, 123-137, 2006.

Chou, Y.J., and Fringer, O.B.: A model for the simulation of coupled flow-bed form evolution in turbulent flows. Journal of Geophysical Research 115, C10041, doi:10.1029/2010JC006103, 2010.

Clifton, H.E.: Wave-formed sedimentary structures: A conceptual model, in Beach and Nearshore Sedimentation, Spec. Publ. 24, edited by R. A. Davis Jr. and R. L. Ethington, pp. 126–148, Soc. for Sediment. Geol., Tulsa, Okla., 1976.

Clifton, H.E., and Dingler, J.R.: Wave-formed structures and paleoenvironmental reconstruction. Marine Geology 60, 165-198, 1984.

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 Sci- ences, 13, 9, 1619–1634, 2009.

Coulibaly, P.: Downscaling daily extreme temperatures with genetic programming. Geophysical Research Letters 31, L16203, doi:10.1029/2004GL020075, 2004.

Cummings, D.I., Dumas, S., Dalrymple, R.W.: Fine-grained versus coarse-grained wave ripples generated experimentally under large-scale oscillatory flow. Journal of Sedimentary Research 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.

Davis, J.P., Walker, D.J., Townsend, M., and Young, I.R.: Wave-formed sediment ripples: Transient analysis of ripple spectral development. Journal of Geophysical Research 109, C07020, doi:10.1029/2004JC002307, 2004.

Delgado Blanco, M., Bell, P., and Montaliu, J.: A new look to the applicability of classical models for ripple prediction. In: Proc. 29th Int. Conf. Coastal Eng. ASCE, Lisbon, Portugal, pp. 1909–1921, 2004.

Diesing, M., Kubicki, A., Winter, C., and Schwarzer, K.: Decadal scale stability of sorted bedforms, German Bight, southeastern North Sea. Cont Shelf Res 26:902–916, 2006.

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

Doucette, J.S.: The distribution of nearshore bedforms and effects on sand suspension on low-energy, micro-tidal beaches in Southwestern Australia. Marine Geology 165, 41-61, 2000.

Doucette, J.S.: Geometry and grain-size sorting of ripples on low-energy sandy beaches: field observations and model predictions. Sedimentology 49, 483-503, 2002.

Doucette, J.S., and O'Donoghue, T.: Response of sand ripples to change in oscillatory flow. Sedimentology 53, 581-596, 2006.

Dumas, S., Arnott, R.W.C., and Southard, J.B.: Experiments on oscillatory-flow and combined-flow bed forms: implications for interpreting parts of the shallow-marine sedimentary record. Journal of Sedimentary Research 75, 501-513, 2005.

Faraci, C., and Foti, E.: Evolution of small scale regular patterns generated by waves propagating over a sandy bottom. Physics of Fluids 13, 1624-1634, 2001.

Faraci, C., and Foti, E.: Geometry, migration and evolution of small-scale bedforms generated by regular and irregular waves. Coastal Engineering 47, 35-52, 2002.

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

Forbes, D.L., and Boyd, R.: Gravel ripples on the inner Scotian Shelf. Journal of Sedimentary Research 57(1) 46-54, 1987.

Foti, E., and Blondeaux, P.: Sea ripple formation: the heterogeneous sediment case. Coastal Engineering 25, 237-253, 1995.

Fredsœ, J., and Deigaard, R.: Mechanics of Coastal Sediment Trans- port, World Sci., Singapore, 1992.

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: Col- let, 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 bed- form 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 Sam- pling Technique for Overfitting Control in Genetic Program- ming, in: EuroGP 2012. LNCS, edited by: Moraglio, A., Silva, S., Krawiec, K., Machado, P., and Cotta, C., Springer, Heidel- berg, 7244, 218–229, 2012.

Grant, W.D., and Madsen, O.S.: Movable bed roughness in unsteady oscillatory flow. Journal of Geophysical Research 87, 469-481, 1982.

Grasmeijer, B.T., and Kleinhans, M.G.: Observed and predicted bed forms and their effect on suspended sand concentrations. Coastal Engineering 51, 351-371, 2004.

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., Black, K.P., 1999. Suspended-sediment reference concentration under waves: field observations and critical analysis of two predictive models. Coastal Engineering 38, 115-141.

Green, M. O. and MacDonald, I. T.: Processes driving estuary in-filling 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.

Hallenbeck, T.R., Kvitek, R.G., and Lindholm, J.: Rippled scour depressions add ecologically significant heterogeneity to soft-bottom habitats on the continental shelf. Marine Ecology Progress Series 468: 119-133, 2012.

Hanes, D.M., Alymov, V., Chang, Y.S., and Jette, C.: Wave-formed sand ripples at Duck, North Carolina. J. Geophys. Res 106, 22575-22592, 2001.

Hansen, J.L., v. Hecke, M., Ellegaard, C., Andersen, K.H., Bohr, T., Haaning, A., and Sams, T.: Stability balloon for two-dimensional vortex ripple patterns. Physical Review Letters 87, 2001.

Harrison, S. E., Locker, S. D., Hine, A. C., Edwards, J. H., Naar, D. F., Twichell, D. C., and Mallinson, D. J.: Sediment-starved sand ridges on a mixed carbonate/siliciclastic inner shelf off west-central Florida, Mar. Geol., 200, 171–194, 2003.

Hay, A.E.: Near-bed turbulence and relict waveformed sand ripples: Observations from the inner shelf. Journal of Geophysical Research 113, C04040, doi:10.1029/2006JC004013, 2008.

Hume, T.M., Green, M.O., and Oldman, J.W.: What happens at the seabed off a headland during a tropical cyclone, Coastal Sediments ’99, Am. Soc. of Civ. Eng., Reston, Va. pp. 1836-1851, 1999.

Hume, T. M., Trembanis, A. C., Hill, A., Liefting, R., and Stephens, S.: Spatially variable, temporally stable, sedimentary facies on an energetic inner shelf, in Coastal Sediments ’03 [CD-ROM], edited by A. Press, Am. Soc. Civ. Eng., Reston, Va, 2003.

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.

Inman, D.L.: Wave-generated ripples in nearshore sands, Tech. Memo. 100, Dep. of the U. S. Army Corps of Eng., Washington, D. C., 1957.

Kambekar, A.R., and Deo, M.C.: Wave prediction using genetic programming and model trees. Journal of Coastal Research 28, 43-50, 2010.

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

Kennedy, J. F., and Falcon, M.: Wave generated sediment ripples, Rep. 86, Hydrodyn. Lab., Mass. Inst. of Technol., Cambridge, 1965.

Khelifa, A., and Ouellet, Y.: Prediction of sand ripple geometry under waves and currents. Journal of waterway, port, coastal, and ocean engineering 126, 14-22, 2000.

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 al- gorithm 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 Com- puters 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.

Lacy, J.R., Rubin, D.M., Ikeda, H., Mokudai, K., and Hanes, D.M.: Bed forms created by simulated waves and currents in a large flume. Journal of Geophysical Research

112, C10018, doi:10.1029/2006JC003942, 2007.

Leckie, D.: Wave-formed, coarse-grained ripples and their relationship to hummocky cross-stratification. Journal of Sedimentary Research, 58, 1988.

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

Li, M.Z., and Amos, C.L.: Sheet flow and large wave ripples under combined waves and currents: field observations, model predictions and effects on boundary layer dynamics. Continental Shelf Research 19, 637-663, 1999.

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

Marieu, V., Bonneton, P., Foster, D.L., and Ardhuin, F.: Modeling of vortex ripple morphodynamics. Journal of Geophysical Research 113, C09007, 2008.

Masselink, G., Austin, M.J., O'Hare, T.J., and Russell, P.E.: Geometry and dynamics of wave ripples in the nearshore zone of a coarse sandy beach. Journal of Geophysical Research 112, C10022, doi:10.1029/2006JC003839, 2007.

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.

McNinch, J.E.: Geologic control in the nearshore: shore-oblique sandbars and shoreline erosional hotspots, Mid-Atlantic Bight, USA. Mar. Geol. 211 (1–2), 121–141, 2004.

Miller, M.C., and Komar, P.D.: A field investigation of the relationship between oscillation ripple spacing and the near-bottom water orbital motions. Journal of Sedimentary Research 50(1) 173-182, 1980a.

Miller, M.C., and Komar, P.D.: Oscillation sand ripples generated by laboratory apparatus. Journal of Sedimentary Research 50(1) 183-191, 1980b.

Mogridge, G.R.: Wave generated bed forms, Ph.D. thesis, Queens Univ., Kingston, Ont., Canada, 1972.

Mogridge, G.R., Davies, M.H., and Willis, D.H.: Geometry prediction for wave-generated bedforms. Coastal Engineering 22, 255-286, 1994.

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.: Some basic concepts of wave sediment transport, Ser. Pap. 20, Inst. Of Hydrodyn. and Hydraul. Eng., Tech Univ. of Den., Lyngby, Denmark, 1979.

Nielsen, P.: Dynamics and geometry of wave-generated ripples. Journal of Geophysical Research 86, 6467-6472, 1981.

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

Nielsen, P.: Coastal bottom boundary layers and sediment transport. World Scientific Publishing Company Incorporated, 1992.

Nishimori, H. and Ouchi, N.: Formation of ripple patterns and dunes by wind-blown sand, Phys. Rev. Lett., 71(1), 197-200, 1993.

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 Kauf- mann, 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.

O'Donoghue, T., and Clubb, G.S.: Sand ripples generated by regular oscillatory flow. Coastal Engineering 44, 101-115, 2001.

O'Donoghue, T., Doucette, J.S., Van der Werf, J.J., and Ribberink, J.S.: The dimensions of sand ripples in full-scale oscillatory flows. Coastal Engineering 53, 997-1012, 2006.

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.

Continental Shelf Research 46, 96-106, 2011.

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. Genetic Programming and Evolvable Machines 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. Journal of Geophysical Research 115, F03043, 2010.

Pedocchi, F., and García, M.H.: Ripple morphology under oscillatory flow: 1. Prediction. Journal of Geophysical Research 114, C12014, doi:10.1029/2009JC005354, 2009a.

Pedocchi, F., and García, M.H.: Ripple morphology under oscillatory flow: 2. Experiments. Journal of Geophysical Research 114, C12015, doi:10.1029/2009JC005356, 2009b.

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

Ribberink, J.S., and Al-Salem, A.A.: Sediment transport in oscillatory boundary layers in cases of rippled beds and sheet flow. Journal of Geophysical Research 99, 12707-12712, 1994.

Ruessink, B. G.: Calibration of nearshore process models: Applica- tion 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 from http://www.eureqa.com/ , 2013.

Silva, S., and Tseng, Y.T.: Classification of Seafloor Habitats Using Genetic

Programming. In: Giacobini, M., Brabazon, A., Cagnoni, S., Di Caro, G.A., Drechsler, R., Ek´art, A., Esparcia-Alc´azar, A.I., Farooq, M., Fink, A., McCormack, J., O'Neill, M., Romero, J., Rothlauf, F., Squillero, G., Uyar, A.S¸., Yang, S. (eds.) EvoWorkshops 2008. LNCS, vol. 4974, pp. 315–324. Springer, Heidelberg, 2008.

Skarke, A., and Trembanis, A.C.: Parameterization of bedform morphology and defect density with fingerprint analysis techniques. Continental Shelf Research 31, 1688-1700, 2011.

Sleath, J.F.A.: The suspension of sand by waves. Journal of Hydraulic Research 20, 439-452, 1982.

Sleath, J.F.A., and Wallbridge, S.: Pickup from rippled beds in oscillatory flow. Journal of Waterway, Port, Coastal, and Ocean Engineering 128, 228-237, 2002.

Smith, J. J., and Wiberg, P. L.: Ripple Geometry in Wave-Dominated Environments Revisited, paper presented at Eos Trans. AGU, 87(36), Ocean Sci. Meet. Suppl., Abstract OS35D-24, 2006.

Soulsby, R.L., and Whitehouse, R.J.S.: Prediction of ripple properties in shelf seas. Mark 2 Predictor for time evolution. Report TR154, HR Wallingford, Wallingford, UK. http://eprints.hrwallingford.co.uk/281/, 2005.

Soulsby, R.L., Whitehouse, R.J.S., and Marten, K.V.: Prediction of time-evolving sand ripples in shelf seas. Continental Shelf Research 38, 47-62, 2012.

Southard, J. B., Lambie, J., Federico, D., Pile, H., and Weidman, C.: Experiments on bed configurations in fine sands under bidirectional purely oscillatory flow, and the origin of hummocky cross-stratification. Journal of Sedimentary Petrology, 60(1), 1–17, 1990.

Styles, R., and Glenn, S. M.: Modeling bottom roughness in the presence of wave-generated ripples, J. Geophys. Res., 107(C8), 3110, doi:10.1029/2001JC000864, 2002.

Thieler, E.R., Pilkey, O.H., Cleary, W.J., and Schwab, W.C.: Modern sedimentation on the shoreface and inner continental shelf at Wrightsville Beach, North Carolina, USA. J Sediment Res 71:958–970, 2001.

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

Thorne, P.D., Williams, J.J., and Davies, A.G.: Suspended sediments under waves measured in a large-scale flume facility. Journal of Geophysical Research, 107(C8), doi:10.1029/2001JC000988, 2002.

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

Traykovski, P.: Observations of wave orbital scale ripples and a nonequilibrium time-dependent model. Journal of Geophysical Research 112, C06026, doi:10.1029/2006JC003811, 2007.

Traykovski, P., Hay, A.E., Irish, J.D., and Lynch, J.F.: Geometry, migration, and evolution of wave orbital ripples at LEO-15. Journal of Geophysical Research 104(C1), 1505–1524, doi:10.1029/1998JC900026, 1999.

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

Trembanis, A.C., and T.M. Hume (2010), Sorted bedforms on the inner shelf off northeastern New Zealand: spatiotemporal relationships and potential paleo-environmental implications, Geo-Mar Letters, DOI 10.1007/s00367-010-0225-8.

van der Werf, J. J., Doucette, J. S., Donoghue, T., and Ribberink, J. S.: Detailed measurements of velocities and suspended sand con- centrations 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 along- shore sediment transport, Nonlin. Processes Geophys., 17, 395– 404, 2010.

Van Oyen, T., de Swart, H. E., and Blondeaux, P.: Bottom topog- raphy 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 Fluid Flow and Surface Waves in Rivers, Estuaries, Seas and Oceans, Aqua, Amsterdam, 1990.

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

van Rijn, L. C.: Principles of Coastal Morphology, Aqua, Amsterdam, 1998.

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 dy- namics, Geology, 27, 727–730, 1999.

Whigham, P.A., Recknagel, F.: Predictive modelling of plankton dynamics in freshwater lakes using genetic programming. In: MODSIM 1999 International Congress on Modelling and Simulation, Modeling and Simulation Society of Australia and New Zealand, 6–9 December, 1999, New Zealand, pp. 691–696, 1999.

Wiberg, P.L., and Harris, C.K.: Ripple geometry in wave-dominated environments. Journal of Geophysical Research 99(C1), 775-775, doi:10.1029/93JC02726, 1994.

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

Wikramanayake, P. N., and Madsen, O. S.: Calculation of movable friction factors, Tech. Rep. DACW-39-88-K-0047, 105 pp., Coastal Eng. Res. Cent., U.S. Army Corps of Eng., Vicksburg, Miss, 1991.

Williams, J.J., Bell, P.S., Thorne, P.D., Trouw, K., Hardcastle, P.J., and Humphery, J.D.: Observed and predicted vertical suspended sediment concentration profiles and bedforms in oscillatory only flow. Journal of Coastal Research, 698-708, 2000.

Williams, J.J., Bell, P.S., and Thorne, P.D.: Unifying large and small wave-generated ripples. Journal of Geophysical Research 110, C02008, doi:10.1029/2004JC002513, 2005.

Williams, J.J., Bell, P.S., Thorne, P.D., Metje, N., and Coates, L.E.: Measurement and prediction of wave-generated suborbital ripples. Journal of Geophysical Research 109, C02004, doi:10.1029/2003JC001882, 2004.

Xu, J.P.: Observations of plan-view sand ripple behavior and spectral wave climate on the inner shelf of San Pedro Bay, California. Continental Shelf Research 25, 373-396, 2005.

Yamaguchi, N., and Sekiguchi, H.: Variability of wave-induced ripple migration in wave-flume experiments and its implications for sediment transport. Coastal Engineering 58, 671-677, 2011.

Yan, B., Zhang, Q.H., and Wai, O.W.H.: Prediction of sand ripple geometry under waves using an artificial neural network. Computers & Geosciences 34, 1655-1664, 2008.

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

You, Z.-J., Yin, B.: A unified criterion for initiation of sediment motion and inception of sheet flow under water waves. Sedimentology 53, 1181-1190, 2006.

Biography

Evan Benjamin Goldstein was born on April 12th 1982 in Santa Monica, California. He graduated from George School in Newtown, PA (2000) followed by Colgate University, NY where he earned a B.A. (Geology) in 2004. Evan started graduate work at Duke University in 2009. He is the first author of 3 articles (the chapters of this dissertation) and received the 'Best Oral Presentation Award' given to a student at the 8th International IAHR symposium on River, Coastal, and Estuarine Morphodynamics in Santander, Spain (June 9-13th, 2013).