EARTH SURFACE PROCESSES AND LANDFORMS
Earth Surf. Process. Landforms (2020)

Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/esp.4811

State of Science

Towards adopting open and data-driven science practices in bed form dynamics research, and some steps to this end

Ronald R. Gutierrez,[1,2]* Alice Lefebvre,[3] Francisco Núñez-González[4] and Humberto Avila[1]
[1] IDEHA, Universidad del Norte, Barranquilla 500001, Colombia
[2] GEOSED, Pontifical Catholic University of Peru, Lima 32, Peru
[3] MARUM, University of Bremen, Bremen 28359, Germany
[4] Leichtweiß-Institut für Wasserbau, University of Braunschweig, Braunschweig 38106, Germany

Received 11 July 2019; Revised 7 January 2020; Accepted 8 January 2020

*Correspondence to: Ronald R. Gutierrez, Universidad del Norte, Km.5 Vía Puerto Colombia, Barranquilla 500001, Colombia. E-mail: rgutierrezll@uninorte.edu.co

ESPL
Earth Surface Processes and Landforms

ABSTRACT: In recent years, open and data-driven science has fostered very important scientific breakthroughs. This study describes the challenges and opportunities for the scientific community devoted to bed form dynamics research in adopting such scientific paradigms through, for example, engineered data sharing, formal recognition of scientists who collect the data, and collaborative development of free accessible software. It highlights that once these actions are completed, the potential application of deep learning techniques could substantially improve bed form models and the scientific understanding of bed form dynamics. Likewise, it discusses the potential of Bedforms-ATM, a free available software, to standardize some bed form data analysis techniques. We propose that the technical challenges be tackled by following scholarly accepted/proposed standards (e.g. FAIR Guiding Principles, Geoscience Papers of the Future), using the body of knowledge being built on the matter by some institutions (e.g. Federation of Earth Science Information Partners), and through technical discussions at scientific meetings such as MARID. © 2020 John Wiley & Sons, Ltd.

KEYWORDS: bed forms; open science; data-driven science; deep learning; scientific software

Introduction

Many rivers and coastal areas around the world are facing increasing demands on both land and water resources for human settlement growth, navigability, and energy. Thus, there is a necessity for improving the prediction of flow and sediment transport for a wide range of rivers and coastal seas. Bed forms are ubiquitous features in shallow and deep-water environments, having a strong influence on flow properties, sediment transport, and biota distribution. They are also of great value for the interpretation of stratigraphic records, which are used as proxies for determining the nature, direction, and strength of palaeoflows (DeCelles *et al.*, 1983; Dasgupta, 2002; Kostic *et al.*, 2019). As such, a better understanding of their dynamics is of relevance for engineers, stratigraphists, geomorphologists, and planers (ASCE, 2002; Best, 2005).

Despite the significant improvement of the scientific understanding of bed form dynamics from field, laboratory, and numerical investigations performed in the last decades, many aspects remain obscure (Allen, 1983; Best, 2005). The emergence of more sophisticated equipment (such as multibeam echo sounders) and data analysis techniques provides a large amount of detailed measurements, which can be analysed to help improve our knowledge on the mutual interaction of bed forms and flow. However, these data may not always be freely available and are certainly not all analysed in a standardized manner. Recently, some progress has been made in this respect: some researchers are starting to share their data, encouraged by the requirements from some funding agencies and scientific journals to make data openly available; many data publishers and repositories are being established; and freely available software for bed form analysis is being developed. For example, Bradley and Venditti (2017) have compiled flow and dune dimension data for bed forms in unidirectional flows, and have made the data available as supplementary material; researchers have published nearly 400 000 datasets to date through the data publisher PANGAEA, including 125 related to bed forms; Gutierrez *et al.* (2018) provided technical details of the free access software named Bedforms-ATM (Bedforms Analysis Toolkit for Multiscale Modeling), which proposes standardizing the scale-based discrimination and dimensionality quantification of natural bed forms in a single platform and whose structure

encourages its expandability via collaboration from the community of users. The use of Bedforms-ATM to conduct scientific research has been reported by Lefebvre (2019).

In recent years, open science and data-driven science have fostered very important scientific discoveries (Hey *et al.*, 2009). Thus, the necessity for the scientific community to move towards an archetype of open and free sharing of scientific data and software has been underlined (McNutt *et al.*, 2016). Some researchers (Peng *et al.*, 2016) even propose that 'universities, research institutions, and funding agencies should develop new measures to evaluate a research project's success not only based on publications and other outcomes it produces but also based on the amount and quality of data it makes available for the wider user community and society'. A similar posture can be found in the EU 'Open Science' initiative and associated report on next-generation responsible metrics (https://ec.europa.eu/research/openscience/pdf/report.pdf). Although some disciplines (e.g. astronomy and oceanography) have a long tradition of sharing data, the community of engineers and scientists devoted to the analysis of river and coastal dynamics, as well as the community dealing with bed form dynamics and stratigraphy, which regularly meets at the River and Marine Dune Dynamics (MARID) conference, is just starting to do so.

The potential of applying machine learning and deep learning techniques over freely available geophysical data to reveal complex physical interactions and processes has latterly been underlined (Reichstein *et al.*, 2019). The plausible application of such techniques to hydrology (Shen, 2018) and turbulence (Duraisamy *et al.*, 2019) has been reported, while Goldstein and Coco (2015) presented the application of machine learning techniques to bed form dynamics. But in spite of the potential of deep learning methods to improve our understanding of bed form dynamics, apparently they have gained little attention. Application of these techniques to bed forms would substantially benefit from data sharing and the standardization of datasets.

The objectives of this contribution are threefold: (1) to discuss the challenges and opportunities of open and free sharing of bed form data based on similar experiences from other disciplines, which could potentially allow the community of fluvial and coastal morphologists and engineers to change its scientific practices in a medium-term time span; (2) to highlight the need to identify ad hoc applications of deep learning techniques to the study of bed form dynamics; and (3) to identify the applications that can be added to Bedforms-ATM that potentially would help improve our understanding of bed form dynamics and/or standardize bed form data analysis in the near future.

The paper is organized as follows: the first three sections elaborate on the first, second, and third objectives of the study, respectively; the fifth section then presents our conclusions.

Changing Our Paradigms

The opportunity

To date, two paradigms are becoming more prevalent in scientific research, namely: (1) openness; and (2) data-intensive scientific approaches, which may even turn into a big-data-intensive scientific approach soon. Openness: (1) enhances the productivity and efficiency of research by preventing the repetition of scientific studies; (2) is essential to the validation of hypotheses, theories, data, and results; and (3) helps to promote trust among scientists by fostering cooperation and collaboration. Openness demands not only allocating information, but also allocating the necessary resources to understand, validate, and apply information such as data, results, methods, and tools (Resnik, 2006; McNutt *et al.*, 2016). Other positive aspects of open science are that it helps socialize science and improve science teaching (Yu *et al.*, 2016).

The data-intensive scientific approach has drawn the attention of the scientific community and has become a new opportunity and incentive for knowledge discovery, because it expands the correlations among multi-disciplinary data, which subsequently triggers the discovery of new models, new rules, and new knowledge (Hey *et al.*, 2009; Guo *et al.*, 2017). This approach has also opened the door to apply deep learning methods to automatically identify complex spatio-temporal patterns, which, after being coupled with physically based models, could substantially improve our understanding of earth system processes (Reichstein *et al.*, 2019).

Currently, monitoring and prediction systems for terrestrial hydrological processes provide information at relatively coarse spatial resolutions (e.g. 5–100 km) over continental to global domains (Wood *et al.*, 2011). However, the necessity and feasibility of developing hyper-resolution data (e.g. 1 km or finer) has been underlined (Wood *et al.*, 2011, 2012), although with some criticisms (Beven and Cloke, 2012). Computational advances in solving hyper-resolution models that will have up to 10^9 unknowns will potentially result in an improved representation of earth processes (Wood *et al.*, 2011). Thus, by coupling such models with bed form models, studying bed form dynamics for large or very large spatial domains could become attainable.

As is the case in many disciplines, we believe that the aforementioned paradigms are not currently fully present in the coastal and river geomorphology community, and thus it represents a pending challenge that could be reached in the coming years through the active participation of community members.

The challenges

Some cultural, institutional, and technological constraints limit openness in many scientific disciplines and possibly, by extension, the evolution towards big-data-driven science. Based on the experience of some of the co-authors of this paper, these limitations are more deeply present in the scientific community from developing countries. Cultural and institutional constraints will possibly be forced to change, as many funding agencies and high-impact-factor peer-reviewed journals have adopted regulations that require scientists to share data, results, methods, and tools (Resnik, 2006; Koutsoyiannis *et al.*, 2016; Resnik *et al.*, 2019), and some countries (e.g. Germany, The Netherlands) pay the fees for their national researchers to publish open access scientific articles in specific journals.

Data are often shared within a working group or with close collaborators before being published and declared to be openly available after publication (Buys and Shaw, 2015). This step of openly sharing data nevertheless is rarely completed. For instance, less than 1% of the collected ecological data is accessible after publication (Reichman *et al.*, 2011), although important discoveries were made by integrating large datasets from many systems, and the potential benefits of data reuse and open data citation are widely recognized (Hey *et al.*, 2009; Lindenmayer and Likens, 2013; Piwowar and Vision, 2013). Some disciplines, such as genomics, have shared repositories, chiefly due to the homogeneity of their data (Reichman *et al.*, 2011).

Past research has underlined the necessity for having access to large amounts of bed form data from well-documented theoretical and experimental case histories, and the need for integrated interdisciplinary studies to fully understand the morphodynamics of bed forms (Allen, 1983; Dalrymple and

Rhodes, 1995; Best, 2005). A similar need was addressed by other scientific communities by starting an open discussion on the necessity, required resources, and schedule to openly share data. After a testing period, some best practices can be adopted and subsequently supported by funding agencies and publication models (McNutt *et al.*, 2016; Yu *et al.*, 2016). For instance, the turbulence-related community started such a discussion in 1980 and published a final report on turbulence models and test cases in 1992 (Bradshaw *et al.*, 1991; Bradshaw, 1992); similarly the limnology community started a discussion in a workshop in 2011 and published a paper on the global lake temperature data they assembled in 2016 (O'Reilly *et al.*, 2016). In our opinion, special sessions to debate on the aforementioned steps should be urgently considered in future MARID meetings. The MARID conference series, which started in 2000 and takes place every 3 to 4 years, provides an ideal environment to discuss bed form research through talks and discussions in a convivial atmosphere with around 80 participants. Since bed form research represents the intellectual interests of geomorphologists, hydraulic engineers, and sedimentologists (Allen, 1976), the conference should encourage the inclusion of representatives from all these disciplines.

Some encouraging steps towards data sharing have been taken recently. Gutierrez *et al.* (2018) reported the development of Bedforms-ATM, a free available software for bed form data analysis, which also provides field bed form data from the Parana River, Argentina. Many authors have made bed form data publicly available. For instance, Gutierrez (2017) published synthetic bed form data which contains ripple, dune, and bar-like features on the information system PANGAEA, which is an Open Access library aimed at archiving, publishing, and distributing georeferenced data from earth system research. Damen *et al.* (2018) identified and analysed the spatial distribution and controlling parameters of sand waves on the Dutch continental shelf. They made their results freely available through 4TU.ResearchData, which provides an archive for long-term access and curation of research datasets, with a focus on data from science, engineering, and technology. Bradley and Venditti (2017) shared the data they used as supplementary information. Bradley and Venditti (2019) deposited their data at university sharing places. Also, some authors specify the repository where the data will be available (Zgheib *et al.*, 2018), but without the precise webpage or without a DOI, so that the data is not easily accessible, even when finally deposited.

As can be seen, published data are placed in a wide variety of scholarly accepted repositories, which in some cases accept data with many formats and place, to some extent, limited requirements on the metadata. This may establish a data ecosystem of very diverse, poorly integrated, and hardly discovered by humans and machines datasets. However, past research suggests that data-driven science requires a program that collects, centrally archives, and documents data and toolboxes under a common initiative that would provide consistency, producibility, and traceability of such information (Durack *et al.*, 2018). To encourage this, for instance, the editors of hydrology journals proposed establishing a jointly agreed protocol for metadata, inspired by a similar initiative in the medical sciences to support the synthesis efforts that build on earlier studies (Koutsoyiannis *et al.*, 2016).

We believe that two major technological challenges will arise in the effort to changing our paradigms, namely: (1) dealing with the heterogeneity of bed form data that results from the lack of common experimental practices, field measurement standards, and data analysis (Reichman *et al.*, 2011; Gutierrez *et al.*, 2013); (2) tracking the provenance of data derived from original datasets, and the scientific outcomes stemming from them through quality control, analysis, and modelling (Reichman *et al.*, 2011). Yet, other aspects may also require special attention as past projects related to the matter suggest that individual communication can be challenging, and that a dedicated data manager to continually maintain and expand the assembled data may be needed (Bradshaw *et al.*, 1991; O'Reilly *et al.*, 2016).

Despite the potential benefits of having access to large bed form data collections, we must be aware that it might induce the proliferation of inductive science (i.e. developing research questions after having data), which, although a valid research method, may contradict the current workflow of science. It also opens the door for the existence of scientists who might hardly be motivated to gather data, because it is time and resource consuming, and who might simply take data gathered by others (Lindenmayer and Likens, 2013; Mueller-Langer and Andreoli-Versbach, 2018). This might be prevented nonetheless by following the good practice that those using open access datasets must work in close collaboration with those who collected these datasets through co-authorship, attribution, or citation (Lindenmayer and Likens, 2013; Chawinga and Zinn, 2019).

Overall, it is recognized that data sharing increases citation rate (Piwowar *et al.*, 2007). Some journals also recognize that this practice has the potential to improve research reproducibility and continuation and submissions' quality (Govindaraju *et al.*, 2019). In this vein, the Geoscience Papers of the Future (GPF) initiative was suggested by Gil *et al.* (2016): it proposes best practices to systematically attain the core concept of reproducibility. Some hydrology publications (Yu *et al.*, 2016) have already reported on their successful application.

We believe likewise that data sharing would advance bed form research in general and profit all those who initially collect and analyse the data, and those who subsequently reuse them.

Tackling the technical challenges

Data need to be stewarded throughout the entire data lifecycle (i.e. from data collection, to management of active datasets, to long-term archiving). However, most disciplines still lack the technical, institutional, and cultural frameworks to support open data access (Peng *et al.*, 2016). Thus, some initiatives for coping with such limitations, such as the FAIR Guiding Principles for scientific data management and stewardship (Wilkinson *et al.*, 2016), have been proposed.

The FAIR Data Principles are intended to set general standards for data towards data-intensive science, and thereby invite that scientific data should be findable (e.g. data are assigned a globally unique and persistent identifier), accessible (e.g. data are retrievable by their identifier using a standardized communications protocol), interoperable (e.g. data use a formal, accessible, shared, and broadly applicable language for knowledge representation), and reusable (e.g. data are adequately described with a plurality of accurate and relevant attributes, and meet the domain-relevant community standards). These principles are expected to systematically facilitate both humans and their machines to discover, get access to, integrate, and analyse scientific data and other scholarly digital objects such as algorithms, code, and workflows that led to published data, among others (Wilkinson *et al.*, 2016). At present, nevertheless, there are concerns over the paucity of their implementation outside the European continent and their relatively limited application in natural science (Van Reisen *et al.*, 2020). This study also reports that there is still important work to do to reach the FAIR tipping point, namely: (1) solving some

problems regarding FAIR implementation (e.g. even in publications that implemented it, there are some metadata issues that limit the downstream use of data); (2) defining policy alternatives (e.g. conceptualization of solutions against poor and inadequate use of data); and (3) identifying political will to address local/regional needs (e.g. the European Commission embraces FAIR application as a public good, while some African nations adopt it for public health initiatives).

The utility of data depends a great deal on appropriate metadata, which needs to adequately describe data content, context, quality, structure, and accessibility to enable searchers to get access to ever-increasing amounts of data in the least amount of time, as well as to be machine actionable (i.e. supplying detailed information to an autonomously acting, computational data explorer) (Fegraus *et al.*, 2005; Wilkinson *et al.*, 2016). Thus, some scientific disciplines have designed ad hoc metadata languages. For instance, the ecology community uses the Ecological Metadata Language (EML) developed by Fegraus *et al.* (2005). EML is implemented in XML (Extensible Markup Language) and has been used by the ecology community to search for data, develop web query tools, integrate heterogeneous datasets, and perform data analysis and visualization (Varadharajan *et al.*, 2019). Data describing bed form datasets possibly require the design of a tailored metadata language and jointly agreed protocols for identifying the provenance and content of the data products.

The Federation of Earth Science Information Partners (ESIP) aims to make earth science data more discoverable, accessible, and usable. In this vein, the ESIP Data Stewardship Committee has provided a set of recommendations, best practices, and guidelines to influence data management carried out by government agencies and other data stewards (Downs *et al.*, 2015). ESIP proposed a provenance and context content standard (PCCS), which lists all content items required to fully represent the provenance and context of the data products resulting from earth science missions, namely: content item name, descriptive definition, rationale (why a given item is needed), criteria (how good the content should be), priority, user community (who would most likely use the item), source, project phase capture, representation (word files, numeric files, etc.), and distribution restrictions (e.g. proprietary concerns). PCCS presented these items in a matrix that is considered a good starting point for developing a standard to offer guidance for data producers, data managers, and others (Ramapriyan *et al.*, 2012). ESIP has stated its openness to apply PCCS standards to other types of data and has encouraged the organizational membership-based participation of earth science data providers (Downs *et al.*, 2015). ESIP has also tested the data stewardship maturity matrix developed by Peng *et al.* (2015), which can be applied by data centres and other data-holding organizations.

We believe that the MARID conferences could be a unique platform to discuss the creation, management, distribution, use, and citation of bed form data, which eventually might lead to devising an organization that can work in partnership with, for instance, ESIP. We are aware that towards this end, funding, a cooperative attitude from the community of engineers and scientists, and collaboration with public and private institutions, and non-governmental organizations, among others, will be necessary.

The bed form dataset that can potentially be built does not necessarily have to be a single system. Instead, it can be made up of centralized multiple crowdsourced open access data entities. Large complex platforms such as Digital Earth are usually built this way (Guo *et al.*, 2017). It is expected that stewarded bed form data would potentially encompass heterogenous, multi-source, multi-temporal, multi-scale, high-dimensional, highly complex, and unstructured geospatial datasets, which are typical characteristics of many geophysical signals datasets (Nativi *et al.*, 2015; Sharma *et al.*, 2015). For instance, due to their specific research interests, stratigraphists (Cornard and Pickering, 2019; West *et al.*, 2019) actively collect and analyse bed form data that may be at different resolution than that used by hydraulic engineers, and compile palaeocurrent trends (Brand *et al.*, 2015). They have also devised schemes to classify cross-strata (Jopling and Walker, 1968; Cheel, 1990) and determine palaeoflows (DeCelles *et al.*, 1983; Dasgupta, 2002), whose products would represent one data entity.

The Potential of Applying Deep Learning Methods

Machine learning is commonly used for clustering, classification, prediction, and pattern recognition operations over a target dataset. Deep learning algorithms are a subset of machine learning algorithms that aim to automatically build sophisticated hierarchical architectures (Reichstein *et al.*, 2019). Most machine learning methods recursively build a model resulting from multiple datasets that are obtained from a target dataset, that is: (1) training data that are used to learn the model; (2) validation data that are used to assess the model fit; and (3) test data that are utilized to evaluate the final model (Flach, 2012).

In the last decade, machine learning techniques such as genetic programming (Goldstein and Coco, 2014; Goldstein *et al.*, 2014) and least-squares support vector machines (Roushangar *et al.*, 2017) have been applied to study bed form dynamics, and the results have been auspicious. For instance, genetic programming, which mimics the evolutive processes (i.e. reproduction and mutation) of biological structures to recursively optimize the size of training data (Goldstein and Coco, 2014), improved the prediction of non-cohesive particle settling velocity and allowed for building a near-bed reference of sediment concentration that captured pattern modes of sorted bed forms that solely deterministic models were not able to describe. Goldstein and Coco (2015) provided a thorough discussion on the combined application of machine learning and deterministic models that describe bed forms.

The use of deep learning techniques in bed form dynamics research has not been thoroughly discussed yet, although ad hoc discussions on its application to other disciplines such as turbulence (Duraisamy *et al.*, 2019) and earth system science (Reichstein *et al.*, 2019) have recently been reported. Some initial steps to the same end in hydrology have also taken place (Trugman *et al.*, 2019). These studies conclude that, although deep learning has been widely used in computer vision, speech recognition, and control systems – which involve processes closely related to those recognized in physics, chemistry, and biology – its application in geosciences is still incipient.

Duraisamy *et al.* (2019) and Reichstein *et al.* (2019) point out that even though deep learning cannot replace physical modelling, it can certainly substantially augment its capabilities by improving model parameterizations, replacing a physical sub-model with a machine learning model, analysing model–observation mismatch, constraining sub-models (i.e. minimizing error propagation of coupled sub-models), and surrogating model or emulation (e.g. emulating parts of full portions of physical models without sacrificing acceptable accuracy). They also underline the fact that successful applications of deep learning will demand: (1) improved interpretability capabilities and physical consistency; (2) feeding it with *a priori* causality relationships; (3) providing it with labelled training data; and (4) satisfying the computational demand that its application

will pose. Similar conclusions were drawn by Goldstein and Coco (2015).

Towards Bed Form Data Analysis Standardization

With more openly accessible bed form data, scientists will require more complex processing and data analysis tools. In this regard, code used to analyse and process data will be a fundamental requirement for transparency and reproducibility (McNutt *et al.*, 2016; Yu *et al.*, 2016). We believe that the Bedforms-ATM platform of Gutierrez *et al.* (2018) could potentially be used to build such code. Bedforms-ATM is currently written in Matlab. To be entirely open access, Bedforms-ATM will probably need moving to an openly available programming language such as Python, which is steadily gaining a leading position for exploratory, interactive, computation-driven scientific research, and scientific data visualization (Millman and Aivazis, 2011; De La Beaujardière, 2019). Likewise, it may need to use standardized units of software (i.e. software containers), to ensure it is portable and run in any operating system.

Bedforms-ATM v1.1 comprises the following applications: (1) bed forms wavelet analysis; (2) power Hovmöller analysis; (3) bed forms multiscale discrimination, which discriminates bed form fields into three scale-based hierarchies (e.g. ripples, dunes, bars); and (4) three-dimensionality analysis, which quantifies the three-dimensionality of bed form fields. Herein we enumerate the applications that can be incorporated into Bedforms-ATM in the coming years.

Decomposition of bed form fields

The decomposition of bed form fields (i.e. the identification and extraction of single bed form entities from bed form fields), and subsequent quantification of its geometric characteristics (e.g. stoss and leeside slopes, wavelengths, and amplitudes), provides information on the interactions of bed morphology and flow field, and sediment transport (Best, 2005). Some researchers (Van der Mark *et al.*, 2008; Gutierrez *et al.*, 2013) have already presented methodologies to perform the decomposition of bed form fields. These methods are easily reproducible or openly available and can therefore be used to standardize the decomposition of bed form fields. They could be incorporated into Bedforms-ATM and be expanded to include, for example, the three-dimensional decomposition methods studied by Ogor (2018).

Bed form statistical analysis

Bed forms decomposition is also necessary for identifying fully developed bed form fields in experimental environments and quantifying the variability of natural bed forms through statistical analysis (Van der Mark *et al.*, 2008; Aberle *et al.*, 2010; Coleman *et al.*, 2011; Gutierrez *et al.*, 2013; Perillo *et al.*, 2014; LeRoy *et al.*, 2016). In his review of river dunes, Best (2005) suggested that large rivers are characterized by bed forms with leeside slope lower than the angle of repose. He further stated that the study of low-angle bed forms constitutes one of the main future research topics in the understanding of bed form dynamics. To this end, sharing data from large rivers worldwide and standardizing bed forms decomposition will be needed. A statistical analysis of bed form fields and their characteristics will enable a better characterization of their properties and provide a possible explanation of their dynamics.

LeRoy *et al.* (2016) reported the development of SABAT (slope-aspect bedform analysis tool), a Matlab tool that does not identify superimposed bed forms but performs a variety of statistical analyses on bed form wavelengths and amplitudes. A similar software was also presented by Cisneros *et al.* (2019). However, to the best of our knowledge, neither such tools are publicly available. Van der Mark *et al.* (2008) quantified the variability in bed form geometry from a series of bathymetric measurements in the lab, in a small river (0.25 m water depth) and the Rhine (8 m water depth). Their analysis could be expanded and deepened in a variety of environments, especially in diverse large and small rivers, in order to statistically describe bed form parameters and their variability.

Bed form classification

The classification of bed forms from natural environments can be performed through the following schemes (Allen, 1976): geometric, where only the bed form morphology is considered and not the flow conditions; and geometric–hydraulic, in which both bed form observations and concurrent hydrodynamic conditions and sediment sizes are available and considered. A challenge of the latter approach is the difficulty of knowing if bed forms are at equilibrium with the flow when measured.

Experimental bed form classification demands a geometric–hydraulic scheme because fully developed bed forms need to be identified. Bedforms-ATM provides a geometric (scale-based) classification of natural bed forms. However, it could be extended to perform a fluvial bed form classification through dimensional and non-dimensional parameters representing flow conditions and sediment properties (e.g. schemes by Liu, Simons and Richardson, Van Rijn), described by Pramono (2005).

In marine environments, bed forms have a great variety of dimensions and shapes, going from small-scale ripples to tidal dunes and sand waves. They are found at a wide range of depths, from the intertidal zone to the continental rise and are subjected to diverse hydrodynamic forcings (e.g. regular and storm waves, and tidal, wave-induced, or contourite currents). They also form in diverse sedimentary settings such as sand, mixed sediment, or sediment starved. Extensive bed form fields are now well known, and the control of environmental parameters on their morphology is better understood (Damen *et al.*, 2018). However, Garlan *et al.* (2016) invites a renewed geometric classification of marine bed forms following the recent improvement of bed form mapping and characterization. This could be done in the framework of a large collaboration bringing marine bed form data together to be analysed in a standardized and comprehensive way. It is likely that this scheme would be performed mainly on a geometric scheme, as concomitant measurement of bathymetry and hydrodynamics is still scarce.

Dumas *et al.* (2005) proposed a geometric–hydraulic classification scheme for experimental bed forms resulting from oscillatory and combined flows. This scheme has also been used successfully to infer the development path of bed forms (Perillo *et al.*, 2014). We believe that after developing a methodology for decomposing bed form fields, the scheme of Dumas *et al.* (2005) could be used to: (1) perform statistical analysis over single entities of bed forms; (2) provided that geometric–hydraulic information is available, study the time it takes for bed forms to become fully developed, which is of special concern for the understanding of bed form morphodynamics (Allen, 1983; Best, 2005; Doré *et al.*, 2016); and (3) better understand the

complex interplay between hydrodynamics, sediment transport, bed form shape, and stratigraphy.

Conclusions

Scientific openness and data-intensive science are becoming more important in today's scientific inquiry. However, they are not currently prevalent in the community of fluvial and coastal morphologists and engineers. We believe that collectively adopting these scientific paradigms will open a myriad of possibilities to better understand the spatio-temporal mechanisms that govern the dynamics and preservation of bed forms, which, at present, remain obscure.

A change towards scientific openness and data-intensive science will pose some technical challenges such as handling large amounts of data, which involves centralizing, transferring, storing, managing, processing, computing, and sharing such data under widely accepted standards – such as those from the FAIR Guiding Principles. Once this is achieved, the use of deep learning techniques could potentially improve the capabilities of physical bed form dynamic models. We believe that the MARID conference series represents a unique platform to discuss the opportunities and challenges related to changing our prevailing scientific practices.

The Bedforms-ATM software can potentially be used as a common platform for standardizing bed form data analysis methodologies via the contribution of river, coastal, and sedimentology engineers and scientists. To improve its capabilities and become freely available *sensu stricto*, it probably needs to migrate to the Python programming language. In our opinion, the following applications could be developed on the Bedforms-ATM platform in the near future: (1) an application to decompose bed form fields (i.e. identifying single bed form entities); (2) an application to perform statistical analysis over single bed form entities; (3) an application to discriminate experimental bed forms and an application to classify marine bed forms.

Acknowledgements—Ronald R. Gutierrez acknowledges Pontificia Universidad Católica del Perú for funding the development of Bedforms-ATM v1.1, and Universidad del Norte for funding this contribution through Project ID 2019-031. Alice Lefebvre is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), Project No. 345915838. We would like to thank the two anonymous reviewers who helped us improve the quality of the manuscript.

Data Availability Statement

Data sharing is not applicable to this paper as no new data were created or analysed in this study.

Conflict of Interest Statement

The authors have no conflict of interest to declare.

References

Aberle J, Nikora V, Henning M, Ettmer B, Hentschel B. 2010. Statistical characterization of bed roughness due to bed forms: a field study in the Elbe River at Aken, Germany. *Water Resources Research* **46**: W03521. https://doi.org/10.1029/2008WR007406.

Allen JR. 1976. Bed forms and unsteady processes: some concepts of classification and response illustrated by common one-way types. *Earth Surface Processes* **1**: 361–374.

Allen JR. 1983. River bedforms: progress and problems. In *Modern and Ancient Fluvial Systems*, Collinson JD, Lewin J (eds). Blackwell: Oxford; 19–33.

ASCE. 2002. Flow and transport over dunes. *Journal of Hydraulic Engineering* **128**: 726–728.

Best J. 2005. The fluid dynamics of river dunes: a review and some future research directions. *Journal of Geophysical Research - Earth Surface* **110**: F04S02. https://doi.org/10.1029/2004JF000218.

Beven KJ, Cloke HL. 2012. Comment on 'Hyperresolution global land surface modeling: meeting a grand challenge for monitoring Earth's terrestrial water' by Eric F. Wood *et al*. *Water Resources Research* **48**: W01801. https://doi.org/10.1029/2011WR010982.

Bradley RW, Venditti JG. 2017. Reevaluating dune scaling relations. *Earth-Science Reviews* **165**: 356–376.

Bradley RW, Venditti JG. 2019. The growth of dunes in rivers. *Journal of Geophysical Research - Earth Surface* **124**: 548–566.

Bradshaw P. 1992. *Final Report on AFOSR 90-0154*. Technical Report, Stanford University.

Bradshaw P, Launder BE, Lumley JL. 1991. Collaborative testing of turbulence models. *ASME Journal of Fluids Engineering* **113**: 3–4.

Brand L, Wang M, Chadwick A. 2015. Global database of paleocurrent trends through the Phanerozoic and Precambrian. *Scientific Data* **2**: 150025. https://doi.org/10.1038/sdata.2015.25

Buys CM, Shaw PL. 2015. Data management practices across an institution: survey and report. *Journal of Librarianship and Scholarly Communication* **3**(2): eP1225. http://doi.org/10.7710/2162-3309.1225

Chawinga WD, Zinn S. 2019. Global perspectives of research data sharing: a systematic literature review. *Library & Information Science Research* **41**: 109–122.

Cheel RJ. 1990. Horizontal lamination and the sequence of bed phases and stratification under upper-flow-regime conditions. *Sedimentology* **37**: 517–529.

Cisneros J, Best J, Dijk T, Mosselman E. 2019. Dune morphology and hysteresis in alluvial channels during long-duration floods revealed using high-temporal resolution MBES bathymetry. *Proceeding of MARID 2019*; 51–56.

Coleman SE, Nikora VI, Aberle J. 2011. Interpretation of alluvial beds through bed-elevation distribution moments. *Water Resources Research* **47**: W11505. https://doi.org/10.1029/2011WR010672.

Cornard PH, Pickering KT. 2019. Supercritical-flow deposits and their distribution in a submarine channel system, Middle Eocene, Ainsa Basin, Spanish Pyrenees. *Journal of Sedimentary Research* **89**: 576–597.

Dalrymple RW, Rhodes RN. 1995. Estuarine dunes and bars. In *Developments in Sedimentology*. Elsevier: Amsterdam; 359–422.

Damen JM, Dijk TA, Hulscher SJ. 2018. Spatially varying environmental properties controlling observed sand wave morphology. *Journal of Geophysical Research - Earth Surface* **123**: 262–280.

Dasgupta P. 2002. Determination of paleocurrent direction from oblique sections of trough cross-stratification – a precise approach. *Journal of Sedimentary Research* **72**: 217–219.

De La Beaujardière J. 2019. A geodata fabric for the 21st century. *Eos* **100**. https://doi.org/10.1029/2019EO136386. Published on 25 November 2019.

DeCelles PG, Langford RP, Schwartz RK. 1983. Two new methods of paleocurrent determination from trough cross-stratification. *Journal of Sedimentary Research* **53**: 629–642.

Doré A, Bonneton P, Marieu V, Garlan T. 2016. Numerical modeling of subaqueous sand dune morphodynamics. *Journal of Geophysical Research - Earth Surface* **121**: 565–587.

Downs RR, Duerr R, Hills DJ, Ramapriyan HK. 2015. Data stewardship in the earth sciences. *D-Lib Magazine* **21**: 7–8. ISSN-e 1082-9873. https://doi.org/10.1045/july2015-downs.

Dumas S, Arnott RW, Southard JB. 2005. 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.

Durack PJ, Taylor KE, Eyring V, Ames SK, Hoang T, Nadeau D, Doutriaux CM, Stockhause M, Gleckler PJ. 2018. Toward standardized data sets for climate model experimentation. *Eos* **99**. https://doi.org/10.1029/2018EO101751. Published on 02 July 2018.

Duraisamy K, Iaccarino G, Xiao H. 2019. Turbulence modeling in the age of data. *Annual Review of Fluid Mechanics* **51**: 357–377.

Fegraus EH, Andelman S, Jones MB, Schildhauer M. 2005. Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America* **86**: 158–168.

Flach P. 2012. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge University Press: Cambridge.

Garlan T, Brenon E, Marches E, Blanpain O. 2016. From regional variability of the morphology of dunes to a new method of their classification. *Proceedings of MARID 2016*; 1–4.

Gil Y, David CH, Demir I, Essawy BT, Fulweiler RW, Goodall JL, Karlstrom L, Lee H, Mills HJ, Oh J-H, Pierce SA, Pope A, Tzeng MW, Villamizar SR, Yu X. 2016. Toward the geoscience paper of the future: best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science* **3**: 388–415.

Goldstein EB, Coco G. 2014. A machine learning approach for the prediction of settling velocity. *Water Resources Research* **50**: 3595–3601.

Goldstein EB, Coco G. 2015. Machine learning components in deterministic models: hybrid synergy in the age of data. *Frontiers in Environmental Science* **3**: 33. https://doi.org/10.3389/fenvs.2015.00033

Goldstein EB, Coco G, Murray AB, Green MO. 2014. Data-driven components in a model of inner-shelf sorted bedforms: a new hybrid model. *Earth Surface Dynamics* **2**: 67–82.

Govindaraju RS, Hantush M, Chu X. 2019. New policy for transparency of data, models, and code. *Journal of Hydrologic Engineering* **24**(3): 01618001.

Guo H, Liu Z, Jiang H, Wang C, Liu J, Liang D. 2017. Big earth data: a new challenge and opportunity for Digital Earth's development. *International Journal of Digital Earth* **10**: 1–12.

Gutierrez RR. 2017. Synthetic data for the Bedforms Analysis Toolkit for Multiscale Modeling (Bedforms-ATM). DOI: https://doi.org/10.1594/PANGAEA.873304.

Gutierrez RR, Abad JD, Parsons DR, Best JL. 2013. Discrimination of bed form scales using robust spline filters and wavelet transforms: methods and application to synthetic signals and bed forms of the Río Paraná, Argentina. *Journal of Geophysical Research - Earth Surface* **118**: 1400–1418.

Gutierrez RR, Mallma JA, Núñez-González F, Link O, Abad JD. 2018. Bedforms-ATM, an open source software to analyze the scale-based hierarchies and dimensionality of natural bed forms. *SoftwareX* **7**: 184–189.

Hey AJ, Tansley S, Tolle KM. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Vol. **1**. Microsoft Research: Redmond, WA.

Jopling AV, Walker RG. 1968. Morphology and origin of ripple-drift cross-lamination, with examples from the Pleistocene of Massachusetts. *Journal of Sedimentary Research* **38**: 971–984.

Kostic S, Casalbore D, Chiocci F, Lang J, Winsemann J. 2019. Role of upper-flow-regime bedforms emplaced by sediment gravity flows in the evolution of deltas. *Journal of Marine Science and Engineering* **7**(1): 5.

Koutsoyiannis D, Blöschl G, Bárdossy A, Cudennec C, Hughes D, Montanari A, Neuweiler I, Savenije H. 2016. Joint editorial: Fostering innovation and improving impact assessment for journal publications in hydrology. *Water Resources Research* **52**: 2399–2402.

Lefebvre A. 2019. Three-dimensional flow above river bedforms: insights from numerical modeling of a natural dune field (Río Paraná, Argentina). *Journal of Geophysical Research - Earth Surface* **124**: 2241–2264.

LeRoy JZ, Rhoads BL, Best JL. 2016. Bed morphology and sedimentology at chute cutoffs: a case study of Mackey Bend, lower Wabash River, IL-IN. *International Conference on Fluvial Hydraulics, RIVER FLOW 2016*; 1736–1742.

Lindenmayer D, Likens GE. 2013. Benchmarking open access science against good science. *Bulletin of the Ecological Society of America* **94**: 338–340.

McNutt M, Lehnert K, Hanson B, Nosek BA, Ellison AM, King JL. 2016. Liberating field science samples and data. *Science* **351**: 1024–1026.

Millman KJ, Aivazis M. 2011. Python for scientists and engineers. *Computing in Science & Engineering* **13**: 9–12.

Mueller-Langer F, Andreoli-Versbach P. 2018. Open access to research data: strategic delay and the ambiguous welfare effects of mandatory data disclosure. *Information Economics and Policy* **42**: 20–34.

Nativi S, Mazzetti P, Santoro M, Papeschi F, Craglia M, Ochiai O. 2015. Big data challenges in building the global earth observation system of systems. *Environmental Modelling & Software* **68**: 1–26.

Ogor J. 2018. *Design of algorithms for the automatic characterization of marine dune morphology and dynamics*. PhD thesis, ENSTA Bretagne.

O'Reilly C, Hampton SE, Sharma S, Gray D, Read JS, Lenters JD, Schneider P. 2016. Challenges in assembling and managing environmental data sets. *Eos* **97**. https://doi.org/10.1029/2018EO044377. Published on 25 January 2016.

Peng G, Privette JL, Kearns EJ, Ritchey NA, Ansari S. 2015. A unified framework for measuring stewardship practices applied to digital environmental datasets. *Data Science Journal* **13**: 231–253.

Peng C, Song X, Jiang H, Zhu Q, Chen H, Chen JM, Gong P, Jie C, Xiang W, Yu G, Zhou X. 2016. Towards a paradigm for open and free sharing of scientific data on global change science in China. *Ecosystem Health and Sustainability* **2**(5): e01225.

Perillo MM, Best JL, Yokokawa M, Sekiguchi T, Takagawa T, Garcia MH. 2014. A unified model for bedform development and equilibrium under unidirectional, oscillatory and combined-flows. *Sedimentology* **61**: 2063–2085.

Piwowar HA, Vision TJ. 2013. Data reuse and the open data citation advantage. *PeerJ* **1**: e175. https://doi.org/10.7717/peerj.175

Piwowar HA, Day RS, Fridsma DB. 2007. Sharing detailed research data is associated with increased citation rate. *PLoS ONE* **2**(3): e308.

Pramono GH. 2005. *The study of bedforms and equivalent roughness sizes in the central Dithmarschen Bight*. PhD dissertation, Kiel University.

Ramapriyan H, Moses J, Duerr R. 2012. Preservation of data for Earth system science – towards a content standard. *2012 IEEE International Geoscience and Remote Sensing Symposium*; 5304–5307.

Reichman OJ, Jones MB, Schildhauer MP. 2011. Challenges and opportunities of open data in ecology. *Science* **331**: 703–705.

Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature* **566**: 195–204.

Resnik DB. 2006. Openness versus secrecy in scientific research. *Episteme* **2**: 135–147.

Resnik DB, Morales M, Landrum R, Shi M, Minnier J, Vasilevsky NA, Champieux RE. 2019. Effect of impact factor and discipline on journal data sharing policies. *Accountability in Research* **26**: 139–156.

Roushangar K, Saghebian SM, Mouaze D. 2017. Predicting characteristics of dune bedforms using PSO-LSSVM. *International Journal of Sediment Research* **32**: 515–526.

Sharma S, Gray DK, Read JS, O'Reilly CM, Schneider P, Qudrat A et al. 2015. A global database of lake surface temperatures collected by in situ and satellite methods from 1985–2009. *Scientific Data* **2**: 150008. https://doi.org/10.1038/sdata.2015.8

Shen C. 2018. Deep learning: a next-generation big-data approach for hydrology. *Eos* **99**. https://doi.org/10.1029/2018EO095649. Published on 25 April 2018.

Trugman DT, Beroza GC, Johnson PA. 2019. Machine learning in geoscience: riding a wave of progress. *Eos* **100**. https://doi.org/10.1029/2019EO122671. Published on 03 May 2019.

Van der Mark CF, Blom A, Hulscher SJ. 2008. Quantification of variability in bedform geometry. *Journal of Geophysical Research - Earth Surface* **113**: F03020. https://doi.org/10.1029/2007JF000940.

Van Reisen M, Stokmans M, Basajja M, Ong'ayo A, Kirkpatrick C, Mons B. 2020. Towards the tipping point of FAIR implementation. *Data Intelligence* **2**: 264–275.

Varadharajan C, Cholia S, Snavely C, Hendrix V, Procopiou C, Swantek D, Riley WJ, Agarwal DA. 2019. Launching an accessible archive of environmental data. *Eos* **100**. https://doi.org/10.1029/2019EO111263. Published on 08 January 2019.

West LM, Perillo MM, Olariu C, Steel RJ. 2019. Multi-event organization of deepwater sediments into bedforms: long-lived, large-scale antidunes preserved in deepwater slopes. *Geology* **47**: 391–394.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**: 160018. https://doi.org/10.1038/sdata.2016.18.

Wood EF, Roundy JK, Troy TJ, Van Beek LP, Bierkens MF, Blyth E et al. 2011. Hyperresolution global land surface modeling: meeting a

grand challenge for monitoring Earth's terrestrial water. *Water Resources Research* **47**(5). https://doi.org/10.1029/2010WR010090.

Wood EF, Roundy JK, Troy TJ, Beek R, Bierkens M, Blyth E et al. 2012. Reply to comment by Keith J. Beven and Hannah L. Cloke on 'Hyperresolution global land surface modeling: meeting a grand challenge for monitoring Earth's terrestrial water'. *Water Resources Research* **48**: W01802. https://doi.org/10.1029/2011WR011202.

Yu X, Duffy CJ, Rousseau AN, Bhatt G, Pardo Álvarez Á, Charron D. 2016. Open science in practice: learning integrated modeling of coupled surface–subsurface flow processes from scratch. *Earth and Space Science* **3**: 190–206.

Zgheib N, Fedele JJ, Hoyal DC, Perillo MM, Balachandar S. 2018. Direct numerical simulation of transverse ripples: 1. Pattern initiation and bedform interactions. *Journal of Geophysical Research - Earth Surface* **123**: 448–477.