Quantifying the quality of coastal morphological predictions

Bosboom, Judith

DOI
10.4233/uuid:e4dc2dfc-6c9c-4849-8aa9-befa3001e2a3

Publication date
2019

Document Version
Final published version

Citation (APA)
Bosboom, J. (2019). *Quantifying the quality of coastal morphological predictions*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:e4dc2dfc-6c9c-4849-8aa9-befa3001e2a3

Quantifying the Quality of Coastal Morphological Predictions

Judith Bosboom

Quantifying the Quality of coastal morphological predictions

Judith Bosboom

Author: Judith Bosboom

Book design by the author—thanks to Sjoerd Leenders for LaTeX help
Cover design by Remco Kalf, Viktorija Usevičiūtė and the author
Figures by the author—thanks to Janbert Aarnink for preparing Figs. 6.1 to 6.3
Printed by Ridderprint BV, Ridderkerk

PhD thesis Delft University of Technology
ISBN 978-94-6384-091-0

An electronic version of this thesis is available at https://repository.tudelft.nl

Quantifying the quality of coastal morphological predictions

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, prof. dr. ir. T.H.J.J. van der Hagen,
Chair of the Board for Doctorates,
to be defended publicly on
Thursday 16 January 2020 at 15:00 o'clock

by **Judith BOSBOOM**
Master of Science in Civil Engineering,
Delft University of Technology, the Netherlands,
born in Nijmegen, the Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee

Rector Magnificus	Chairperson
Prof. dr. ir. A.J.H.M. Reniers	Delft University of Technology, promotor
Prof. dr. ir. M.J.F. Stive	Delft University of Technology, promotor

Independent members

Dr. M.A. Davidson	University of Plymouth, United Kingdom
Prof. dr. S.J.M.H. Hulscher	University of Twente
Prof. dr. B.G. Ruessink	University of Utrecht
Prof. dr. ir. A.W. Heemink	Delft University of Technology

Reserve member

Prof. dr. ir. S.G.J. Aarninkhof	Delft University of Technology

Other member

Dr. ir. C.F. de Valk	Royal Netherlands Meteorological Institute

This research was financially supported by Delft University of Technology. Travel and conference costs were covered by NWO/ASPASIA grant (015.007.051).

Contents

Summary

The quality of morphodynamic predictions is often indicated by a skill score that measures the relative accuracy of a morphological prediction over a prediction of zero morphological change, using the mean-squared error (MSE) as the accuracy measure. Through a generic classification based on skill levels, predictions receive a quality label. As simple as this Brier skill score (BSS) or $\mathrm{MSESS}_{\mathrm{ini}}$[1] may seem, it is not well understood and, hence, sometimes misinterpreted. Further, as a point-wise accuracy metric, the MSE heavily penalizes small misplacements of coastal features such as scour holes, bars or channels—a phenomenon referred to as the "double penalty effect". From the perspective of a coastal morphologist, this may lead to wrong decisions as to which of two predictions is better. Motivated by the above, this thesis investigates the behaviour of the $\mathrm{MSESS}_{\mathrm{ini}}$ as well as explores and develops validation methods and corresponding error metrics that, as opposed to point-wise metrics, take the spatial structure of morphological patterns into account.

Formulations and classifications for $\mathrm{MSESS}_{\mathrm{ini}}$—with and without accounting for measurement error—are examined by using synthetic examples, examples from literature and a long-yearly Delft3D model simulation compared to measurements. It is shown that the common reference of zero change fails to make model performance comparable across different prediction situations (geographical locations, forcing conditions, time periods, internal dynamics). Also, it is demonstrated that the combined presence of larger, persistent scales and smaller, intermittent scales in the cumulative bed changes may lead to an apparent increase of skill with time, without the prediction on either of these scales becoming more skilful with time. Further, the $\mathrm{MSESS}_{\mathrm{ini}}$ is shown to have the tendency to favour model results that underestimate the variance of cumulative bed changes, a feature inherited from the MSE. As a consequence of these limitations, the $\mathrm{MSESS}_{\mathrm{ini}}$ may report a relative ranking of predictions not matching the intuitive judgement of experts. Further, it is shown theoretically and through an artificial case of rip channel formation, that the existing methods to correct for measurement error are inconsistent in either their skill formulation or their suggested classification scheme.

In order to overcome the inherent limitations of point-wise metrics, three novel diagnostic tools for the spatial validation of 2D morphological predictions are developed. The first method deforms the predictions towards the observations, minimizing the point-wise squared error. Error measures are then formulated based on both the smooth displacement field between predictions and observations and the residual point-wise error field after the deformation. This field deformation

[1] Mean-squared-error skill score with the initial bed—denoted with the subscript "ini"—as the reference prediction or a reference prediction of zero change.

method is shown to outperform the convential approach based on the point-wise root-mean-squared error (RMSE) for a variety of morphological fields—generated with Delft3D—for an idealized case of a tidal inlet. Since it optimizes the location of individual depth values by (locally) stretching or compressing the predicted morphological pattern, the method is seen to capture the visual closeness of morphological patterns. Sediment mass continuity, however, is not guaranteed.

The second method defines the distance between predicted and observed morphological fields in terms of an optimal sediment transport field that moves the misplaced sediment from the predicted to the observed morphology. The optimal corrective transport field has the lowest quadratic transportation cost and is relatively easily found by solving an elliptic partial differential equation. The root-mean-squared value of the optimal transport field—the root-mean-squared transport error (RMSTE)—is proposed as a new error metric. It is put to the test for simple 1D and 2D cases as well as for the more realistic morphological fields of the above mentioned schematized tidal inlet. The results show that the RMSTE, as opposed to the RMSE, is able to discriminate between predictions that differ in the misplacement distance of predicted morphological features, and avoids the consistent favouring of the underprediction of morphological variability that the RMSE is prone to. As opposed to the field deformation method, the optimal transport method is mass-conserving, parameter-free and symmetric.

The third method is a scale-selective validation approach that allows any metric to selectively address multiple spatial scales. It employs a smoothing filter in such a way that—in addition to the domain-averaged statistics—localized validation statistics and maps of prediction quality are obtained per scale. The term "scale" as considered by this method refers to geographic extent or areal size of focus. The employed skill score weights how well the morphological structure and variability are simulated, while avoiding the double penalty effect by which point-wise accuracy metrics tend to reward the underestimation of variability. The scale-selective method is demonstrated by application to measured and computed bathymetric fields.

Finally, it is recommended that a combination of metrics is used in the validation of morphological models and that the weighting is determined by the goal of the simulation. In such a set of metrics, point-wise metrics should be supplemented with an error decomposition, as to avoid undesired underestimation of variability. Further, a set of performance metrics must include a metric—e.g. the RMSTE—that accounts for the spatial structure of the observed and predicted morphological fields. In future studies, the behaviour of the RMSTE in a range of practical applications needs to be considered. In order to do so, an extension of its implementation to arbitrary model domains is required. It may also be worthwile, albeit nontrivial, to explore possibilities to solve the optimization problem with a linear instead of with a quadratic cost function.

Samenvatting

De kwaliteit van morfodynamische voorspellingen wordt vaak aangegeven met een *skill score* die de relatieve nauwkeurigheid van een morfologische voorspelling meet ten opzichte van een voorspelling zonder morfologische veranderingen, waarbij de gemiddelde kwadratische fout of *mean-squared error* (MSE) als foutmaat wordt gebruikt. Via een generieke classificatie op basis van skillwaarden krijgen voorspellingen een kwaliteitslabel. Deze Brier skill score (BSS) of $\text{MSESS}_{\text{ini}}$[1] is minder eenvoudig dan hij wellicht lijkt en wordt daarom soms verkeerd geïnterpreteerd. Omdat metingen en berekeningen puntsgewijs worden vergeleken, bestraft de MSE kleine positiefouten van morfologische fenomenen zoals erosiekuilen, zandbanken en geulen relatief zwaar, een fenomeen dat bekend staat als het "*double penalty effect*". Vanuit het perspectief van een kustmorfoloog, kan dit leiden tot verkeerde beslissingen ten aanzien van welke van twee voorspellingen beter is. Ingegeven door het bovenstaande, onderzoekt dit proefschrift het gedrag van de $\text{MSESS}_{\text{ini}}$ en verkent en ontwikkelt het validatiemethoden en bijbehorende foutmaten die, in tegenstelling tot puntsgewijze foutmaten, de ruimtelijke structuur van morfologische patronen in beschouwing nemen.

Formuleringen en classificaties voor $\text{MSESS}_{\text{ini}}$, met en zonder correcties voor meetfouten, worden onderzocht aan de hand van kunstmatige voorbeelden, voorbeelden uit de literatuur en een langjarige Delft3D modelsimulatie, die vergeleken wordt met meetresultaten. Er wordt aangetoond dat het gebruikelijke referentiemodel (waarin er geen morfologische verandering optreedt) de kwaliteit van voorspellingen niet vergelijkbaar weet te maken voor uiteenlopende voorspellingssituaties (geografische locaties, forcering, tijdsperioden, interne dynamiek). Ook wordt gedemonstreerd dat de gecombineerde aanwezigheid van grotere, persistente schalen en kleinere, intermitterende schalen in de cumulatieve bodemveranderingen kan leiden tot een schijnbare toename van skill met de tijd, zonder dat voor (een van) deze schalen afzonderlijk de skill daadwerkelijk toeneemt. Het onderzoek wijst verder uit dat de $\text{MSESS}_{\text{ini}}$ de neiging heeft om de voorkeur te geven aan modelresultaten die de variantie van cumulatieve bodemveranderingen onderschatten, een eigenschap die wordt doorgegeven door de MSE. Als gevolg van deze beperkingen, is het mogelijk dat de $\text{MSESS}_{\text{ini}}$ een rangorde van voorspellingen rapporteert die niet overeenkomt met het intuïtieve oordeel van experts. Verder wordt aangetoond, op basis van theoretische overwegingen en een kunstmatige casus van muivorming, dat de bestaande methoden om te corrigeren voor meetfouten inconsistent zijn in ofwel hun skillformulering ofwel hun voorgestelde classificatieschema.

[1] Skill score gebaseerd op de MSE waarbij het subscript "ini" verwijst naar de initiële bodem als de referentievoorspelling, ofwel een referentievoorspelling zonder morfologische veranderingen.

Om een oplossing te vinden voor de inherente beperkingen van puntsgewijze foutmaten, zijn drie innovatieve diagnostische methoden ontwikkeld voor de ruimtelijke validatie van 2D morfologische voorspellingen. De eerste methode vervormt de voorspellingen in de richting van de waarnemingen om zo de puntsgewijze kwadratische fout te minimaliseren. Foutmaten worden vervolgens gebaseerd op het gladde verplaatsingsveld tussen voorspellingen en waarnemingen en op het resterende puntsgewijze foutveld na de vervorming. Deze veldvervormingsmethode blijkt voor een verscheidenheid aan (met Delft3D gegenereerde) morfologische velden voor een geïdealiseerd getijdenbekken beter te presteren dan de conventionele aanpak op basis van de puntsgewijze *root-mean-squared error* (RMSE, de vierkantswortel uit de MSE). Doordat de locatie van individuele dieptewaarden geoptimaliseerd wordt door het voorspelde morfologische patroon (lokaal) uit te rekken of te comprimeren, is deze methode in staat om de visuele nabijheid van morfologische patronen vast te leggen. Massabehoud van het sediment is echter niet gegarandeerd.

De tweede methode definieert de afstand tussen voorspelde en waargenomen morfologische velden in termen van een optimaal sedimenttransportveld dat het verkeerd gepositioneerde sediment van de voorspelde naar de waargenomen morfologie beweegt. Het optimale corrigerende transportveld heeft de laagste kwadratische transportkosten en is relatief eenvoudig te vinden door een elliptische partiële differentiaalvergelijking op te lossen. De vierkantswortel van het gemiddelde kwadratische optimale transport, ofwel de *root-mean-squared transport error* (RMSTE), wordt voorgesteld als een nieuwe foutmaat. Deze wordt getest voor eenvoudige 1D- en 2D-voorbeelden, evenals voor de meer realistische morfologische velden van het bovengenoemde schematische getijdenbekken. De resultaten laten zien dat de RMSTE, in tegenstelling tot de RMSE, in staat is onderscheid te maken tussen voorspellingen die verschillen in de mate waarin voorspelde morfologische kenmerken verkeerd gepositioneerd zijn. Ook vermijdt de RMSTE de consistente voorkeur voor onderschatting van morfologische variabiliteit, waar de RMSE de neiging toe heeft. De optimale transportmethode is massabehoudend en symmetrisch en kent geen parameters, in tegenstelling tot de veldvervormingsmethode.

De derde methode is een schaalselectieve validatiemethode die het een willekeurige prestatiemaat mogelijk maakt om selectief meerdere ruimtelijke schalen te adresseren. De methode maakt daarbij gebruik van een ruimtelijk filter, op een zodanige manier dat, naast de domeingemiddelde statistieken, ook gelokaliseerde validatiestatistieken en ruimtelijke velden van voorspellingskwaliteit per schaal worden verkregen. De term “schaal”, zoals die in deze methode wordt gebruikt, verwijst naar de geografische omvang of grootte van het aandachtsgebied. De gebruikte skill score weegt hoe goed de morfologische structuur en variabiliteit worden gesimuleerd, waarbij het double penalty effect waardoor puntsgewijze foutmaten vaak de onderschatting van variabiliteit belonen, wordt vermeden. De schaalselectieve methode wordt gedemonstreerd door deze toe te passen op gemeten en berekende bathymetrische velden.

Ten slotte wordt aanbevolen dat een combinatie van prestatiematen wordt gebruikt bij de validatie van morfologische modellen en dat de weging hiervan wordt bepaald door het doel van de simulatie. In een dergelijke set van prestatiematen worden puntsgewijze foutmaten bij voorkeur aangevuld met een ontleding van de fout, om zo ongewenste onderschatting van variabiliteit te voorkomen. Verder dient een set prestatiematen een maat zoals de RMSTE te bevatten, die de ruimtelijke structuur van de waargenomen en voorspelde morfologische velden in beschouwing neemt. In toekomstige studies zal het gedrag van de RMSTE in een reeks praktische toepassingen moeten worden onderzocht. Om dit te kunnen doen, is een uitbreiding van de implementatie van de RMSTE naar willekeurige modeldomeinen vereist. Het kan ook de moeite waard zijn, hoewel verre van triviaal, om de mogelijkheden te verkennen om het optimalisatieprobleem op te lossen met een lineaire in plaats van met een kwadratische kostenfunctie.

. .

1 Introduction

Coastal morphological predictions typically are the 2D-gridded outcomes of coastal area models consisting of bed levels at high resolution. This thesis is about quantifying the quality of such predictions, which is an essential part of both calibration and validation of morphodynamic models. Calibration is the common engineering practice to adjust the modelling parameters so that improved agreement with the experimental data is obtained, whereas validation is the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model (Oberkampf and Trucano, 2002). The quantification of the agreement between computational results and experimental data assumes that some appropriate measure of correspondence can be established—a performance or validation metric. As every statistical measure condenses a large number of data into a single value, it only provides one projection of the model errors emphasizing a certain aspect of the error characteristics of the model performance (Chai and Draxler, 2014). Various metrics are thus required to adequately represent the enormous amount of information—bed levels for each grid point and the complex relationships between the grid points—contained in morphological fields. The selection of the metrics and their weighting should be driven by application requirements, with as primary consideration what the model must predict in conjunction with what data is available (Thacker et al., 2004b).

This introductory chapter is structured as follows. First, Sect. 1.1 describes the current validation practice of morphological fields, which relies heavily on grid-point based accuracy and skill metrics. Grid-point based accuracy metrics measure the averaged correspondence between individual pairs of model outcomes and observations, whereas corresponding skill metrics determine the accuracy relative to the accuracy of a prediction produced by a standard of reference (Murphy, 1993). Next, Sect. 1.2 examines strategies for the development of innovative performance metrics that, as opposed to point-wise metrics, are able to account for the spatial interdependency of the observed and predicted fields. Finally, the objectives and outline of this thesis are elaborated on in Sect. 1.3 and Sect. 1.4, respectively.

1.1 The MSE and BSS in morphodynamic model validation

The oldest method for evaluating the quality of 2D morphological predictions is by eye-ball comparison of patterns of sedimentation and erosion between observations and simulations. The power of this qualitative validation technique lies in the fact that the human brain is incredibly good at identifying patterns. The visual inspection requires looking through the eye and filtering the output to identify position, magnitude and orientation of certain features of interest and using human judgement to discern the prediction errors. This makes eye-ball or visual validation prone to individual and subjective biases of interpretation. Besides, it is increasingly difficult to apply if there are multiple predictions (as in a sensitivity analysis or ensemble prediction).

Quantitative validation methods are often grid-point based; they compare observations and predictions per grid point and compute various metrics for the entire set or subset of grid points. Gallagher et al. (1998) and Sutherland et al. (2004) introduced the concept of skill to morphodynamic model validation. A skill score measures the relative accuracy of the prediction over some reference prediction. For a prediction with accuracy E, a skill score can be formulated as follows:

$$\mathrm{ESS} = \frac{E - E_r}{E_\mathrm{i} - E_r} \tag{1.1}$$

where E_r is the accuracy of a baseline or reference prediction and E_i the accuracy of an impeccable (perfect) prediction. The ESS ranges from $-\infty$ to 1, with negative (positive) values indicating a prediction worse (better) than the reference prediction. A value between 0 and 1 can be interpreted as the proportion of improvement over the reference prediction.

For deterministic predictions of continuous variables, such as seabed elevation, a common choice for the accuracy measure E in Eq. 1.1 is the mean-squared error (MSE). The resulting skill score is often referred to as mean-squared-error skill score outside our field, e.g. Murphy (1988), but is named Brier skill score (BSS) by coastal modellers following Sutherland et al. (2004)[1]. It reads:

$$\mathrm{MSESS} = \frac{\mathrm{MSE} - \mathrm{MSE}_r}{0 - \mathrm{MSE}_r} = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE}_r} \tag{1.2}$$

since the MSE of a perfect prediction $\mathrm{MSE}_\mathrm{i} = 0$. In morphodynamic modelling, it is common practice to use the initial observed bathymetry at the start of a simulation as the reference. The $\mathrm{MSESS}_\mathrm{ini}$—the MSESS with the initial bed as the reference—can be considered as the fraction of improvement of the model results compared to a model that predicts that no morphodynamic change will occur. It is valued through a generic classification for morphodynamic computations, which distinguishes between bad, poor, reasonable, good and excellent predictions depending on the skill value (Sutherland et al., 2004). Through the Murphey–Epstein decomposition of the MSE into phase, amplitude, and map-mean error, the $\mathrm{MSESS}_\mathrm{ini}$ can be decomposed into various error components (Murphy and Epstein, 1989; Sutherland et al., 2004).

[1] This thesis addresses this skill metric for *nonprobabilistic* variables as mean-squared-error skill score (MSESS), consistent with Murphy (1988). Technically, the term Brier skill score (BSS) is reserved for the relative accuracy of *probabilistic forecasts* with the Brier score (Brier, 1950) as the accuracy measure, which is a mean-squared error for *probabilistic* forecasts with two mutually-exclusive outcomes (e.g. rain or no rain).

The $\mathrm{MSESS}_{\mathrm{ini}}$ a.k.a. the BSS has quickly become widely accepted amongst morphodynamic modellers as the preferred way of demonstrating model skill. Conclusions about (relative) model performance and model sensitivities are not seldom largely based on the $\mathrm{MSESS}_{\mathrm{ini}}$ (see references in amongst others Sects. 2.1 and 5.1). Nonetheless, little attention has been paid to the interpretation of the $\mathrm{MSESS}_{\mathrm{ini}}$ and its values. Also, the Murphy–Epstein decomposition, which may provide additional insight into the aspects of prediction quality measured by the $\mathrm{MSESS}_{\mathrm{ini}}$, is not often used. Consequently, the use of the $\mathrm{MSESS}_{\mathrm{ini}}$ carries the risk of an implicit redefinition of quality through optimizing its scores, especially when used as the single validation metric and in automated calibration procedures.

In order to account for measurement error, adjusted MSESS formulations and skill classifications have been suggested by van Rijn et al. (2003) and Sutherland et al. (2004). Unfortunately, this has initiated an inconsistent use of skill definitions and rankings in subsequent literature. Therefore, the establishment of the best method to take measurement error into account is called for.

For the $\mathrm{MSESS}_{\mathrm{ini}}$ to allow the intercomparison of quality across a range of prediction situations, the zero change model must correctly reflect the intrinsic difficulty of prediction situations (Winkler, 1994; Murphy, 1988; Wilks, 2011; Brier and Allen, 1951) with a different morphological development prior to the evaluation time—for instance trend-wise, cyclic or episodic. Since the accuracy of the zero change model is given by the observed cumulative morphological development away from the initial bed, the $\mathrm{MSESS}_{\mathrm{ini}}$ normalizes the error in the bed levels by the observed cumulative change. Therefore, it can be expected that the stringency of the skill test depends on the state of the initial morhology, for instance whether the chosen initial morphology is pre-storm or post-storm or whether simulations are initialized from a smooth or a high-variability initial bottom. Similarly, the $\mathrm{MSESS}_{\mathrm{ini}}$ can be expected to develop differently in time for a trend than for a seasonal system, due to the difference between gross and net change. This raises the question whether the $\mathrm{MSESS}_{\mathrm{ini}}$ can create the “level playing field” (Winkler et al., 1996) required for an intercomparison of skill values.

Whether the $\mathrm{MSESS}_{\mathrm{ini}}$ is the appropriate metric, given what a morphodynamic model must predict, further depends on the characteristics of the MSE. There is a consensus amongst morphologists that the generally high-variability predictions[2] of high-resolution models are useful if they can reproduce features such as scour holes and bar or channel generation and migration, even with small space and timing errors. Unfortunately, as a point-wise accuracy metric, the MSE tends to penalize, rather than reward, the model’s capability to provide information on these features of interest, a phenomenon also referred to as the “double penalty

[2] In general, the term high-variability prediction may refer to predictions exhibiting short-scale variability in space and/or time. In this thesis, the term variability is mostly used to refer to the spatial variability of bed levels or sedimentation and erosion patterns, measured by the standard deviation or variance at the scales of interest.

effect" (see Sect. 4.1); inevitable location (and timing) errors in a high-variability prediction will lead to a larger MSE than for smoother predictions of, for instance, a lower-resolution model (Bougeault, 2003; van Rijn et al., 2003). It is therefore difficult to demonstrate the quality of a high-variability morphodynamic prediction.

1.2 Methods for spatial validation of coastal morphology

The validation of the small-scale morphological variability, usually found in high-resolution coastal morphological predictions, brings about a range of new validation questions. Are there spatial displacement errors? Is the variability well represented at all scales? Is it necessary to accurately predict shorter-scale features to make reliable longer-term predictions? At which spatial scales does the model have sufficient skill? Does the skill vary within the model domain? These questions are not easily addressed with the traditional validation approach. First, any single-number metric suffers from considerable loss of information. Moreover, the essential quantities of interest in the patterns of morphology and morphological change are not captured by point-wise validation metrics, such as the MSE and MSESS. Indeed, point-wise metrics tend to penalize rather than reward the prediction of features if these features are somewhat displaced in space (Sect. 1.1). Clearly, there is a need for alternative validation methods that account for spatial information contained in predicted and observed fields.

The need for spatial validation methods also stems from possible limits to practical predictability of morphological change. Accuracy measures or skill scores are inappropriate when the small scales are unpredictable because the information on those scales can be regarded as noise. However, a prediction with little skill on small scales may still be useful over a larger area (e.g. an ebb-tidal delta). Also, amplitude, shape and spacing of rhythmic features like sand bars may be predicted reasonably well, although a deterministic location may not be predictable. Filtering high-resolution details by eye-ball validation implicitly acknowledges that the practical predictability on small scales is limited, but may be better on the larger scales of certain features of interest. Nevertheless, model output is mostly presented at the scale of the computational grid. This may cause untrained users of predictions to overestimate the model credibility on small spatial scales. On the other hand, when comparing predictions and observations side-by-side the presence of information on unskilful scales may also lead to a false sense of model failure.

In response to the undesirable properties of traditional point-wise metrics when applied to high-resolution predictions, researchers in various fields, amongst others meteorology, have proposed numerous new methods to assess the model performance, the majority of which can be grouped into two categories (Gilleland et al., 2009, 2010a): filtering methods and deformation methods. Filtering methods apply a spatial filter to the predicted and observed fields or to the difference field, and then calculate overall statistics on the filtered fields to evaluate performance at various scales. Applied filters are either smoothing filters or bandpass spatial filters (Fourier, wavelets, etc.). Deformation methods deform predicted features or fields in order to obtain a better match with the observations and determine error statistics based on the required spatial manipulation (displacement, rotations, scaling, etc.) and the residual errors after manipulation.

The ideas behind the deformation and filtering methods provide useful starting points for the development of dedicated spatial validation methods for coastal morphology. With our usual 2D coastal morhological predictions in mind, a deformation method that directly targets fields is more practical than and, thus, preferable over a feature-based approach. By (locally) stretching or compressing the morphological pattern, a typical field deformation method would optimize the location of pixels with given predicted intensities (depth values) in order to achieve a better match with the observations. An advantage of such pattern matching by shifting image pixels is that it may be relatively close to the visual validation by morphologists. On the other hand, it could be disadvantageous that sediment is not necessarily conserved, since pixels rather than sand are moved. Sediment conservation would be guaranteed if the optimal transformation from predictions to observations is defined in terms of the physical quantity responsible for morphodynamic development: sediment transport. The quest for such a transformation would bring us to the mathematical domain of optimal mass transport, which deals with the transport of a distribution of mass to another distribution of mass on the same space, in such a way as to keep the transportation cost to a minimum (Santambrogio, 2015; Villani, 2003). The transformation of predictions towards the observations, whether by image matching or optimal transport, must then be supplemented by the formulation of appropriate error metrics based on it.

Filtering approaches have the advantage of selectively addressing multiple scales of interest in the morphology or sedimentation/erosion patterns. For 2D morphology and arbitrarily shaped model domains, however, the application of band-pass filters is far from trivial and the physical interpretation of the results is difficult, since the scales are not easily linked to morphological features. Methods based on smoothing filters—also called neighbourhood methods—are appealing due to their simplicity of operation and interpretation; a filter is applied at progressively coarser scales, yielding progressively smoother fields, and summary statistics are applied to the filtered fields. Common smoothing methods, however, are limited in the aspects of model performance that can be considered. For instance, no information on spatial variation of performance in the model domain is provided. A useful validation framework for coastal morphology would employ a smoothing filter in such a way that, in addition to domain-averaged statistics, *localized* validation statistics are obtained. This could be achieved by the computation of validation statistics in a sliding window, similar to localized data analysis (Fotheringham et al., 2002). Appropriate validation statistics should take both similarity in structure and amplitude of the patterns into account, while avoiding the double penalty problem (Sect. 1.1).

1.3 Approach

The overarching aim of this thesis is to contribute to an improved validation assessment of morphological predictions, in particular field predictions. It pursues two main research objectives, which derive from Sects. 1.1 and 1.2, respectively. These two objectives are formulated and elaborated in research questions and objectives as follows:

Objective 1 Investigate the behaviour of the commonly used $\text{MSESS}_{\text{ini}}$ (Eq. 1.1 with the initial bed as the reference prediction) a.k.a. the Brier skill score (BSS). This first objective is addressed in Chs. 2 and 3. Research questions are:

1.1. What is the effect on the $\text{MSESS}_{\text{ini}}$ of the use of the point-wise mean-squared error (MSE) as the accuracy measure? (Chs. 2 and 3)

1.2. What is the added value and correct interpretation of the Murphy–Epstein decomposition of the $\text{MSESS}_{\text{ini}}$? (Chs. 2 and 3)

1.3. What is the rationale behind taking measurement error into account and how should this translate to skill formulations and rankings? (Ch. 3)

1.4. To what extent does the zero change model underlying the $\text{MSESS}_{\text{ini}}$ make model performance comparable across different prediction situations (geographical locations, forcing conditions, time periods, internal dynamics)? (Chs. 2 and 3)

Objective 2 Develop validation methods and corresponding performance metrics that take the spatial structure of morphological patterns into account. This second objective is addressed in Chs. 4 to 6. Specific research objectives and questions are:

2.1. Develop a field deformation method suited for the validation of morphological patterns and formulate (an) appropriate error metric(s) to be used in conjunction with this method. (Ch. 4)

2.2. What is the behaviour of the error metric(s) as referred to in Objective 2.1, in comparison to the behaviour of point-wise metrics? (Ch. 4)

2.3. Develop an optimal transport method for the validation of morphological patterns and derive (a) corresponding error metric(s). (Ch. 5)

2.4. What is the behaviour of the error metric(s) as referred to in Objective 2.3, in comparison to the behaviour of point-wise metrics? (Ch. 5)

2.5. Develop a scale-selective validation framework that resolves the spatial distribution of appropriate validation statistics for multiple scales. (Ch. 6)

2.6. What information is provided by the scale-selective framework as mentioned in Objective 2.5 and what is the added value of addressing multiple scales? (Ch. 6)

1.4 Thesis outline

The core of this thesis consists of four published papers (Chs. 2 to 4 and 6) and one manuscript that is currently under review (Ch. 5). Even though the respective chapters can therefore be read independently, they are strongly related. In order to clarify their interrelationship as well as provide a quick overview of the highlights, a brief introduction to the paper is given at the start of each chapter.

Chapters 2 and 3 pursue Objective 1 by evaluating the current validation practice of morphological fields and particularly the MSE-based skill metric with the zero change model as the reference (the $\mathrm{MSESS}_{\mathrm{ini}}$ a.k.a. the BSS). Next, in Ch. 4, new error metrics are introduced based on an image matching or warping method, which finds the smooth displacement field between predictions and observations that minimizes the point-wise error (Objective 2.1 and Question 2.2). Chapter 5 then presents a diagnostic tool—including a novel error metric—that moves misplaced sediment from the predicted to the observed morphology through an optimal, rotation-free sediment transport field (Objective 2.3 and Question 2.4). Subsequently, Objective 2.5 and Question 2.6 are addressed in Ch. 6, which introduces a scale-selective validation method for 2D morphological predictions that provides information on the variation of model skill with spatial scale and within the model domain.

Finally, Ch. 7 is a concluding chapter providing a comprehensive overview of the findings of this thesis as well as discussing recommendations for further research.

2 On the perception of morphodynamic model skill

This chapter is republished with minor changes only from J. Bosboom, A.J.H.M. Reniers and A.P. Luijendijk (2014). On the perception of morphodynamic model skill. *Coastal Engineering 94*, pp. 112–125, doi:10.1016/j.coastaleng.2014.08.008.

It explores the behaviour of the mean-squared-error skill score (MSESS) a.k.a. the Brier skill score (BSS), which is a widely used metric to evaluate and classify the performance of morphological models. Nonetheless, surprisingly little is known about which aspects of quality are exactly measured by the BSS. Also, the premise that its values can be used to compare predictions across different prediction situations—geographical locations, forcing conditions, time periods, internal dynamics—has not been critically evaluated. This chapter, in conjunction with Ch. 3 (i.e. Bosboom and Reniers, 2018), attempts to fill these gaps. The highlights of Ch. 2 are:

1. Synthetic examples, an example from literature and a long-yearly Delft 3D simulation are used to evaluate the BSS.
2. Visual inspection by experts leads to a different perception of skill than the BSS.
3. In the presence of inevitable location errors, the BSS favours predictions that underestimate the variance of the bed changes.
4. The normalization with the cumulative bed change, which stems from the initial bed as the reference, is not able to create a “level playing field”.
5. An increase in skill with time can result from the emerging of the more skilful larger scales, without the skill on these scales increasing in time.
6. A generic ranking, based on BSS values, has limited validity.
7. Multiple performance metrics are required in order to fully describe prediction quality.

Abstract

The quality of morphodynamic predictions is generally expressed by an overall grid-point based skill score, which measures the relative accuracy of a morphological prediction over a prediction of zero morphological change, using the mean-squared error (MSE) as the accuracy measure. Through a generic ranking for morphodynamic model predictions, this MSE-based skill score (MSESS) aims at making model performance comparable across different prediction situations (geographical locations, forcing conditions, time periods, internal dynamics). The implicit assumptions underlying this approach are that the MSE is an appropriate measure of correspondence for morphological predictions and that the accuracy of the initial bed as the reference correctly reflects the inherent difficulty or ease of prediction situations. This paper presents a thorough analysis of the perception of model skill through the MSE skill score. Using synthetic examples, an example from literature and a long-yearly Delft3D model simulation, we demonstrate that unexpected skill may be reported due to a violation of either of the above assumptions. It is shown that the accuracy of the reference fails to reflect the relative difficulty of prediction situations with a different morphological development prior to the evaluation time (for instance trend, cyclic/seasonal, episodic, speed of the development). We further demonstrate that the MSESS tends to favour model results that underestimate the variance of cumulative bed changes, a feature inherited from the MSE. As a consequence of these limitations, the MSESS may report a relative ranking of predictions not matching the intuitive judgement of experts. Guidelines are suggested for how to adjust calibration and validation procedures to be more in line with a morphologist's expert judgement.

2.1 Introduction

A commonly-used, single-number metric for judging the relative accuracy of morphodynamic simulations is the mean-squared-error skill score (MSESS) that goes by the name Brier skill score (BSS)[1] among morphodynamic modellers (Sutherland et al., 2004). It measures the proportion of improvement in accuracy of a prediction over a reference model prediction, using the mean-squared error (MSE) as the accuracy measure. Generally, the initial bed is chosen as the reference prediction, which implies a reference model of zero morphological change. To our knowledge, Gallagher et al. (1998) were the first to determine morphodynamic model skill as the model accuracy relative to the accuracy of the initial bathymetry. They used the root-mean-squared error (RMSE) as the accuracy measure. Several other researchers and modellers have determined the MSESS with the measured initial bathymetry as the reference for field and laboratory applications of both cross-shore profile models (e.g. van Rijn et al., 2003; Sutherland et al., 2004; Henderson et al., 2004; Pedrozo-Acuña et al., 2006; Ruessink et al., 2007; Roelvink et al., 2009; Ruggiero et al., 2009; Walstra et al., 2012; Williams et al., 2012) and area models (e.g. Sutherland et al., 2004; Scott and Mason, 2007; McCall et al., 2010; Ganju et al., 2011; Orzech et al., 2011; van der Wegen et al., 2011; Dam et al., 2013; Fortunato et al., 2014). The simulation duration for the field cases varied from days for bar evolution to decades for large-scale tidal basin evolution. Alongside MSESS, its decomposition according to Murphy and Epstein (1989) has been used to separately

[1] We prefer to address this skill metric as MSESS, consistent with Murphy (1988). Technically, the term Brier skill score (BSS) is reserved for the relative accuracy of probabilistic forecasts with the Brier score (Brier, 1950) as the accuracy measure, which is a mean-squared error for probabilistic forecasts with two mutually-exclusive outcomes (e.g. rain or no rain).

assess phase and amplitude errors (Sutherland et al., 2004; Ruessink and Kuriyama, 2008; van der Wegen et al., 2011; van der Wegen and Roelvink, 2012).

Values for the MSESS are typically computed for the entire spatial array at a particular time and valued through a generic ranking for morphodynamic computations (van Rijn et al., 2003; Sutherland et al., 2004). This approach, which aims at making model performance comparable across different prediction situations (geographical locations, forcing conditions, time periods, internal dynamics) has become the standard in quantitative judgement of morphodynamic model skill (Roelvink and Reniers, 2012). Gallagher et al. (1998) already pointed out that a comparative analysis based on skill values requires a good understanding of the statistics of predictive skill. Nonetheless, the behaviour of MSESS and the validity of a generic ranking based on its values have not been thoroughly explored. Also, there have been accounts of skill scores not matching the researcher's perception of model performance. For instance, van der Wegen and Roelvink (2012) suggested that their relatively high skill scores were a result of the use of a horizontally uniform initial bed (and hence of a low accuracy of the reference model). For bed profile predictions, Walstra et al. (2012) reported skill values to increase in time to an unexpectedly similar level as previously found for weekly timescales by Ruessink et al. (2007).

Clearly, a crucial element of skill is the proper selection of the reference; it establishes the zero point at the scale on which skill is measured and, hence, defines a minimal level of acceptable performance. Therefore, a comparative analysis based on skill scores is only effective to the extent that the intrinsic difficulty of different prediction situations is correctly reflected in the level of accuracy of the reference predictions (Brier and Allen, 1951; Winkler, 1994; Murphy, 1988; Wilks, 2011). In weather forecasting, where skill scores have widely been used for over a century (Murphy, 1996a), the reference is generally required to be an unskilful, yet not unreasonable forecast as can be made with a naive forecasting method (Winkler, 1994). Examples are persistence, i.e. the observations at a given time are forecast to persist, and long-term climatology, i.e. the average of historical data is used as the baseline (Murphy, 1996b). The naive method that produces the most accurate forecasts is considered the appropriate method in a particular context (Murphy, 1992). Hence, for short-term weather forecasts, persistence is generally the more appropriate choice of reference, whereas climatology may be better for longer-term predictions. The reference of zero morphological change is similar to the concept of persistence in that it assumes the morphology to persist, i.e. remain unchanged, in time. However, instead of using a recent state (e.g. the previously observed value) as the reference, as is common practice in weather forecasting, the zero change model is applied irrespective of the prediction horizon, by assuming the *initial* bed to persist. Another marked difference is the cumulative nature of morphology as the persisted parameter, as opposed to for instance precipitation. Thus, the accuracy of the zero change model is given by the observed cumulative morphological development away from the initial bed, which must adequately represent the situation's inherent difficulty for the MSESS to create a "level playing field" (Winkler et al., 1996).

Not only the choice of reference, but also the choice of the accuracy measure determines the reported skill. Unfortunately, grid-point based accuracy measures, such as the MSE, are prone to reward predictions that underestimate variability (Anthes, 1983; Taylor, 2001; Mass et al., 2002), a phenomenon also referred to as the "double penalty effect" (Bougeault, 2003). As a consequence, such accuracy measures may lead to wrong decisions as to which of two morphological predictions is better (Bosboom and Reniers, 2014b, i.e. Ch. 4). If this undesirable property is inherited by the MSESS, the diagnosis of model skill will similarly be affected.

The purpose of this paper is to investigate the potential impact of the choice of the zero change reference model, in combination with the MSE as the accuracy measure, on the perception of morphodynamic model skill. First, Sect. 2.2 provides a review and discussion on the interpretation of the conventional skill metrics used in morphodynamic skill assessment, viz. the MSESS and its Murphy–Epstein decomposition. It includes examples, both synthetic and from literature, which demonstrate how unexpected skill can be obtained by using the MSESS. Next, in Sect. 2.3, a record of bathymetric data and Delft3D morphodynamic computations, spanning 15 years, is used to illustrate that also for a real-life case, the common skill metrics may lead to an interpretation of model performance inconsistent with expert judgement. In Sect. 2.4, the implications for morphological model validation are discussed. Finally, Sect. 2.5 presents conclusions and discusses avenues for adaptation of validation strategies.

2.2 A critical review of the common skill metrics

This section reviews the skill metrics as commonly applied for morphodynamic model validation. Possible pitfalls for the perception of model performance are identified and illustrated with various examples. First, Sect. 2.2.1 summarizes the MSESS and its Murphy–Epstein decomposition (Murphy and Epstein, 1989) for arbitrary spatial fields and a yet undefined reference. Second, in Sect. 2.2.2, the metrics are interpreted in the context of the validation of morphological fields, using the initial bed as the reference. Third, Sect. 2.2.3 discusses the impact of the zero change reference model on the perception of morphodynamic model skill. Finally, Sect. 2.2.4 demonstrates that the MSESS tends to reward an underestimation of the variance of bed changes.

2.2.1 Mean-squared-error skill score

The concept of skill, according to Murphy (1996a) first proposed by Gilbert (1884), refers to the relative accuracy of a prediction over some reference or baseline prediction. For a prediction with accuracy E, a generic skill score ESS with respect to a reference prediction with accuracy E_r is (e.g. Sutherland et al., 2004):

$$\mathrm{ESS} = \frac{E - E_r}{E_\mathrm{i} - E_r} \tag{2.1}$$

where E_i is the accuracy of an impeccable prediction. A prediction that is as good as the reference prediction receives a score of 0 and an impeccable prediction a score of 1. A value between 0 and 1 can be interpreted as the proportion of improvement over the reference prediction. If the MSE is used as the accuracy measure, Eq. 2.1 yields (Murphy, 1988):

$$\mathrm{MSESS} = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE}_r} \tag{2.2}$$

since $\mathrm{MSE_i} = 0$. The MSESS ranges from $-\infty$ to 1, with negative (positive) values indicating a prediction worse (better) than the reference prediction.

The MSE between the predicted and observed spatial fields is defined as:

$$\mathrm{MSE} = \left\langle \left(p - o\right)^2 \right\rangle = \frac{1}{n} \sum_i^n w_i \left(p_i - o_i\right)^2 \tag{2.3}$$

where the angle brackets denote spatially weighted averaging, (p_i, o_i) are the ith pair of the gridded predicted and observed fields p and o respectively and n is the number of points in the spatial domain. Further, w_i is a weighting factor by grid-cell size, such that $\sum_i^n w_i = n$ and for regularly spaced grids $w_i = 1$.

Skill metrics often are in terms of the differences (anomalies) with respect to the reference prediction r. With the anomalies of predictions and observations given by $p' = p - r$ and $o' = o - r$, respectively, we can rewrite Eq. 2.3 upon substitution as:

$$\mathrm{MSE} = \left\langle \left(p' - o'\right)^2 \right\rangle. \tag{2.4}$$

Further, the accuracy of the reference prediction is given by:

$$\mathrm{MSE}_r = \left\langle \left(r - o\right)^2 \right\rangle = \left\langle o'^2 \right\rangle. \tag{2.5}$$

An advantage of the mean-squared-error measure of accuracy and the corresponding MSESS is that they can readily be decomposed into components that describe specific elements of prediction quality. The decomposition according to Murphy and Epstein (1989) separates the MSE into correlation and conditional and systematic bias terms (Appendix 2.A). Herewith, Eq. 2.4 can be written as (cf. Eqs. 2.14 and 2.15):

$$\mathrm{MSE} = \sigma_{o'}^2 \left(1 - \alpha' + \beta' + \gamma'\right) \tag{2.6}$$

with

$$\alpha' = \rho^2_{p'o'} \tag{2.7a}$$

$$\beta' = \left(\rho_{p'o'} - \frac{\sigma_{p'}}{\sigma_{o'}}\right)^2 \tag{2.7b}$$

$$\gamma' = \frac{\left(\overline{p'} - \overline{o'}\right)^2}{\sigma^2_{o'}}. \tag{2.7c}$$

Here $\overline{p'}$ and $\overline{o'}$ are the weighted map means and $\sigma_{p'}$ and $\sigma_{o'}$ are the weighted standard deviations of p' and o'. Further, $\rho_{p'o'} = \sigma_{p'o'}/(\sigma_{p'}\sigma_{o'})$ is the weighted Pearson correlation coefficient between p' and o', with $\sigma_{p'o'}$ representing the weighted covariance. Note that the MSE can be considered as the summation of $\mathrm{MSE}_{\mathrm{bias}} = \sigma^2_{o'}\gamma'$ that expresses the systematic bias or map-mean error and $\mathrm{MSE}_{\mathrm{fluct}} = \sigma^2_{o'}(1 - \alpha' + \beta')$ that quantifies the mismatch between the fluctuating parts in predictions and observations.

Equivalently, we can write for MSE_r:

$$\mathrm{MSE}_r = \sigma^2_{o'}\left(1 + \epsilon'\right) \tag{2.8}$$

where

$$\epsilon' = \frac{\overline{o'}^{\,2}}{\sigma^2_{o'}} \tag{2.9}$$

is nonzero if the map mean of the observations differs from the map mean of the reference prediction.

Finally, substitution of Eqs. 2.6 and 2.8 in Eq. 2.2 yields the Murphy–Epstein decomposition of the skill score (Murphy and Epstein, 1989):

$$\mathrm{MSESS} = \frac{\alpha' - \beta' - \gamma' + \epsilon'}{1 + \epsilon'}. \tag{2.10}$$

Livezey et al. (1995) explained $1 - \alpha'$ as the phase error and α' as the phase association between predicted and observed anomalies, β' as a penalty due to conditional bias or amplitude error of the anomalies (with a penalty for both insufficient and excessive predicted amplitudes) and γ' as the reduction of skill due to map-mean errors. Hence, α' can be regarded as the skill in the absence of biases.

2.2.2 Reference model of zero morphological change

In morphodynamic modelling, the predictand is the bathymetry, such that p and o in Eq. 2.3 are the predicted and observed bed levels z_p and z_o, respectively. In order to determine the relative accuracy of bed level predictions, it is a common practice to use the initial observed bathymetry at the start of the simulation as the reference prediction, which implies that the model to beat is a model of zero morphological change. In that case, the anomalies are the cumulative sedimentation/erosion fields from the simulation start time $t = 0$: $p' = \Delta z_p$ and $o' = \Delta z_o$. Herewith, from Eqs. 2.3 to 2.5 we have $\mathrm{MSE} = \langle (z_p - z_o)^2 \rangle = \langle (\Delta z_p - \Delta z_o)^2 \rangle$ and $\mathrm{MSE}_r = \langle \Delta z_o^2 \rangle$. Upon substitution, Eq. 2.2 leads to a skill score valid for the zero change reference model:

$$\mathrm{MSESS}_{\mathrm{ini}} = 1 - \frac{\left\langle \left(\Delta z_p - \Delta z_o\right)^2 \right\rangle}{\left\langle \Delta z_o^2 \right\rangle} \tag{2.11}$$

with the angle brackets again indicating spatially weighted averaging.

The $\mathrm{MSESS}_{\mathrm{ini}}$ expresses the proportion of improvement in the accuracy of bed level predictions or, equivalently, of predictions of cumulative sedimentation/erosion over a model that predicts no morphological change. It is often interpreted as the model added accuracy relative to a situation in which no modelling is done (although technically the zero change model is a model as well, albeit a naive one). The proportion of improvement is typically valued through a generic ranking for morphodynamic computations (van Rijn et al., 2003; Sutherland et al., 2004). Table 2.1 shows the ranking proposed by Sutherland et al. (2004) for the skill formulation according to Eq. 2.11. Note that slightly different rankings have been proposed in combination with skill formulations that include observation error (van Rijn et al., 2003; Sutherland et al., 2004).

	$\mathrm{MSESS}_{\mathrm{ini}}$
Excellent	1.0−0.5
Good	0.5−0.2
Reasonable/fair	0.2−0.1
Poor	0.1−0.0
Bad	<0.0

Table 2.1: Classification according to Sutherland et al. (2004) for the MSE skill score as in Eq. 2.11.

With the anomalies equal to the cumulative sedimentation/erosion fields, Eqs. 2.7 and 2.9 can be written as $\alpha' = \rho^2_{\Delta z_p \Delta z_o}$, $\beta' = (\rho_{\Delta z_p \Delta z_o} - \sigma_{\Delta z_p}/\sigma_{\Delta z_o})^2$, $\gamma' = (\overline{\Delta z_p} - \overline{\Delta z_o})^2/\sigma^2_{\Delta z_o}$ and $\epsilon' = \overline{\Delta z_o}^{\,2}/\sigma^2_{\Delta z_o}$. For the normalization term ϵ', nonzero values are obtained in the case of an observed net sediment import or export from the initial time to the evaluation time (Gerritsen et al., 2011). A nonzero γ' indicates a misestimation of the amount of sediment that has been imported into or exported from the model domain and, equivalently, of the mean bed levels. Hence, γ' can be considered as a (normalized) sediment budget error (Gerritsen et al., 2011). Following Livezey et al. (1995), Sutherland et al. (2004) refer to $1 - \alpha'$ and β' as measures of phase and amplitude errors, respectively, of the cumulative sedimentation/erosion fields (see Sect. 2.2.1). Note that the phase and amplitude errors of predicted *bed levels* are given by $1 - \alpha$ and β (Eqs. 2.15a and 2.15b) rather than $1 - \alpha'$ and β'. Only in the special case that the reference prediction is a horizontal bed (e.g. van der Wegen and Roelvink, 2012), we have $\alpha' = \alpha$, $\beta' = \beta$ and $\gamma' = \gamma$.

The phase error $1 - \alpha'$ is often loosely interpreted as a position error, signifying that "sand has been moved to the wrong *position*" (Sutherland et al., 2004). Gerritsen et al. (2011) explain the phase association α' as the degree of similarity between the spatial patterns of sedimentation and erosion. Since the correlation coefficient measures the tendency of the predictions and observations to vary together (Appendix 2.A), a nonperfect phase association ($\alpha' < 1$) may result from incorrect locations, shapes and relative magnitudes of the sedimentation/erosion features. Predictions that are different by a constant or a constant proportion (either positive or negative) receive the same α'. Therefore, we prefer to consider α' as the extent to which the *structure* of the predicted and observed sedimentation/erosion fields is similar and recognize that overall *magnitudes* of predicted and observed bed changes may not be close for $\alpha' = 1$. With α' measuring the structural similarity, its complement $1 - \alpha'$ measures the structural dissimilarity between the predicted and observed sedimentation/erosion fields.

According to Sutherland et al. (2004), a nonzero amplitude error β' indicates that "the wrong *volumes* of sand have been moved", whereas Gerritsen et al. (2011) refer to β' as a transport rate error. Section 2.2.4 demonstrates that these interpretations should be used with care, but first the impact of the zero change reference model on the perception of model skill is discussed.

2.2.3 Morphodynamic model skill as (mis)perceived using the zero change model

In Eq. 2.11, the MSE is normalized with MSE_r and hence with the observed mean-squared cumulative bed changes $\langle \Delta z_o^2 \rangle$. This means that for the zero change model to be an adequate reference model enabling cross-comparison and absolute ranking of predictions, the net bed changes from the start time of the simulations must represent an evaluator's judgements about the difficulty of predictions for different situations and simulation times. In this section, we reason that this requirement cannot be expected to hold and that consequently the perception of model skill may be distorted.

Let us first consider two hypothetical regions characterized by an identical, propagating morphological feature. During the considered time period, both features have moved over the same net distance, such that the net displaced sediment volumes are equal. However, one feature has propagated at a steady speed to its final position, while the other feature has first moved in the opposite direction under the influence of an episodic event, and subsequently slowly moved back, under milder conditions, to its final position. Although the latter situation would generally be considered the more difficult prediction situation, cumulative (net) changes cannot discern between the two.

As a second example, we consider a cross-shore profile development with a summer–winter cycle and small, random variations between the same seasons in consecutive years. Now, a cross-shore profile model is initialized from a profile measured in winter and run for several years, covering a number of winter–summer profile cycles. For all consecutive modelled winter profiles, the accuracy of the reference is high, such that a similar, high accuracy is required to obtain a certain level of skill. For the modelled summer profiles on the contrary, each summer a similar, lesser accuracy is required, since the initial winter bed is not a good estimate for the observed summer profile. Given a constant modelled accuracy, the diagnosed temporal evolution of model skill would therefore show an artificial seasonal trend with higher skill in summer, but with no changes between the same seasons from year to year.

The above examples demonstrate that observed cumulative bed changes are not likely to be a proper indicator of the inherent ease or difficulty of a morphological prediction, since they do not reflect the nature of the morphological development prior to the evaluation time, but only its cumulative effect. The $\mathrm{MSESS_{ini}}$ could thus very well make the wrong decision as to which of two predictions is better, by awarding a higher skill based merely on a lower accuracy of the initial bed as the reference and not through any intrinsic higher prediction skill. Consequently, the validity of judging morphodynamic model performance based on $\mathrm{MSESS_{ini}}$, through a ranking as in Table 2.1, may be less generic than often assumed. Note that in weather forecasting, this complication is not encountered in the same manner, since predictands such as precipitation, as opposed to morphology, are not cumulative. Also, persistence of the initial situation is only used for a short enough lag, i.e. as long as persistence can still be considered a reasonable prediction (e.g. at the scale of days for short-range forecasts).

For longer-range simulations of seasonal systems, a more appropriate naive prediction could be the initial or last observed state for the same season (e.g. "next July is like this July", hence a one-year persistence model). By using a one-year persistence model for inter-seasonal modelling of seasonal morphodynamics, artificial seasonal variation of skill due to the varying accuracy of the reference can be avoided. The zero change model may only provide a fair reference as long as the model-data comparison is performed yearly, at the same phase in the seasonal cycle as the initial bed.

Still, even if the zero change reference model is only applied yearly, values of $\mathrm{MSESS_{ini}}$ for a long-yearly simulation of a seasonal system and an equally long simulation of a progressive development should not be compared. For the progressive development, the use of the zero change reference model implies that in time, the minimal level of acceptable performance is lowered at a rate determined by the cumulative (net) observed bed changes. Of course, it could be argued that the progressive lowering of the (metaphorical) bar qualitatively agrees with a modeller's intuition that it is only fair that for a longer time in the simulation, and hence a more difficult prediction situation, a lesser accuracy is required to achieve a certain skill level. This interpretation, however, is not consistent with the fact that the zero change reference model for seasonal systems does not exhibit a similar relaxation of the stringency of the test over the course of multiple years, regardless of the amount of gross change. As a consequence, the simulation of the trend has an unfair advantage over the simulation of the seasonal system and increasingly so further into the simulation.

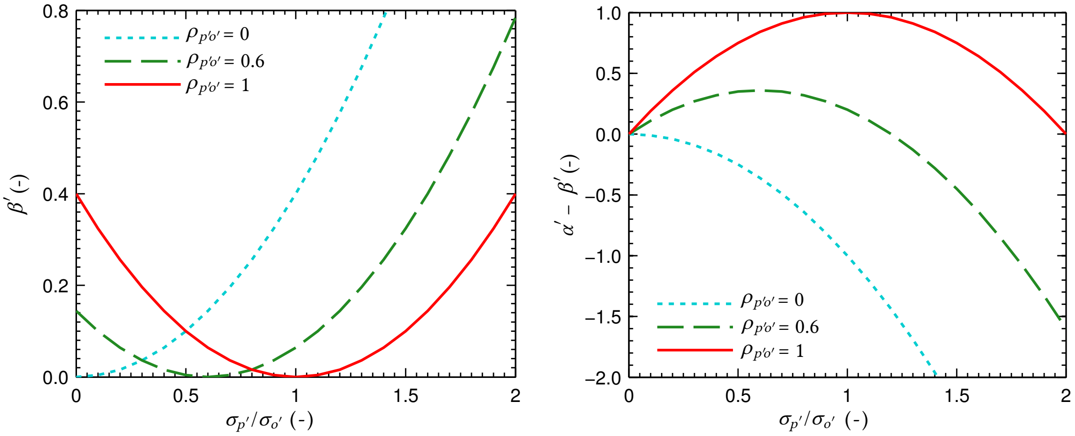
In conclusion, observed mean-squared cumulative bed changes cannot be expected to accurately reflect and thus effectively neutralize the level of difficulty among different prediction situations and times in a simulation. This places severe limits on the general validity of a comparative analysis based on $\mathrm{MSESS}_{\mathrm{ini}}$. On a case-by-case basis, $\mathrm{MSESS}_{\mathrm{ini}}$, notably its time-evolution for a trend, may still provide useful information. Therefore, Sect. 2.3 thoroughly investigates how to interpret the temporal variation of $\mathrm{MSESS}_{\mathrm{ini}}$ for a real-life case that shows a consistent bathymetric development away from the initial bed.

2.2.4 Underestimation of the variance of bed changes through the use of $\mathrm{MSESS}_{\mathrm{ini}}$

In this section, we demonstrate that $\mathrm{MSESS}_{\mathrm{ini}}$ is prone to reward predictions that underestimate the overall magnitude of bed changes. To this end, we analyze the Murphy–Epstein decomposition of $\mathrm{MSESS}_{\mathrm{ini}}$, notably the amplitude error β'.

The behaviour of β', which is controlled by $\sigma_{p'}/\sigma_{o'}$ and $\rho_{p'o'}$ (Eq. 2.7b), is shown in Fig. 2.1a for $\rho_{p'o'} = 0, 0.6$ and 1. The line for $\rho_{p'o'} = 0.6$ is characteristic of the behaviour of β' for a suboptimal correlation, for instance a situation of an erosion hole that is slightly misplaced, such that $0 < \rho_{p'o'} < 1$; even if the erosion hole is predicted correctly with respect to size ($\sigma_{p'} = \sigma_{o'}$), the amplitude error β' is nonzero. In fact, the amplitude error β' is minimized for $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'}$. As a result, the interpretation of a nonzero β' reflecting that the wrong volumes of sand have been moved is only strictly valid for $\rho_{p'o'} = 1$ (Sutherland et al., 2004).

The above also implies that for positive correlation, the skill score $\mathrm{MSESS}_{\mathrm{ini}}$ is maximized for $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'}$ (Eq. 2.10 and Fig. 2.1b). This shows an undesirable property of the MSE skill score, namely that for the same suboptimal anomaly correlation, a higher skill would have been reported for $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'}$ than for $\sigma_{p'}/\sigma_{o'} = 1$, such that sedimentation/erosion fields that underpredict the overall amount of sedimentation and erosion may be favoured above predictions with the correct variance of the bed changes. As can be seen from Eq. 2.15b, this feature is



(a) (b)

Figure 2.1: Amplitude error β' and skill score MSESS$_{\text{ini}}$ = $\alpha' - \beta'$ (assuming $\gamma' = \epsilon' = 0$ in Eq. 2.10) versus $\sigma_{p'}/\sigma_{o'}$ for $\rho_{p'o'}$ equal to 0, 0.6 and 1: **(a)** β' has a minimum for $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'}$, **(b)** the skill $\alpha' - \beta'$ is maximized for $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'}$.

inherited from the MSE, which is known for its tendency to reward the underestimation of the variability (e.g. Arpe et al., 1985; Gupta et al., 2009; Bosboom and Reniers, 2014b, i.e. Ch. 4).

Interestingly, a real-life illustration is provided by the comparison of observed and predicted bathymetric changes for East Pole Sand, reported in Sutherland et al. (2004). Since three predictions, which only differ with respect to the values of the representative grain diameter, are compared for the same prediction situation (their Fig. 4) and hence relative to the same initial bed, the ranking between them is not affected by the normalization with the accuracy of the reference. Also, the values of ϵ' are equal. From their Fig. 4 and Table 9, it can be seen that among the three predictions that have the same positive, but nonperfect correlation between predicted and measured bed changes ($\rho_{p'o'} = \sqrt{0.38} = 0.62$), the MSESS$_{\text{ini}}$ favours the prediction for which $\sigma_{p'}/\sigma_{o'}$ is the closest to $\rho_{p'o'}$ (and thus β' is the smallest, viz. $\beta' = 0.01$). The values of γ' are small and do not differ significantly for the three predictions. As a result, the prediction with the coarsest grain size, for which the standard deviation of the bed changes deviates most from the observations ($\sigma_{p'}/\sigma_{o'} = 0.52$ or 0.72, cf. Fig. 2.1a[2]), is diagnosed with the highest skill (MSESS$_{\text{ini}}$ = 0.34, 0.29, 0.15 for D_{50} = 0.5, 0.35, 0.25 mm, respectively). It is likely however, that an expert, asked to visually compare the quality of these sedimentation/erosion fields, would not prefer this prediction, as for the coarsest grain size the (maximum) magnitudes of sedimentation and erosion are clearly

[2] For D_{50} = 0.5 mm, we have $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'} - \sqrt{\beta'} = 0.62 \pm 0.1$. Observing from their Fig. 4 that $\sigma_{p'}/\sigma_{o'}$ increases with decreasing grain size, we deduce, using the values in their Table 9, that for D_{50} = 0.35 mm and 0.25 mm, $\sigma_{p'}/\sigma_{o'}$ = 0.88 and 1.07, respectively.

underestimated. Apparently, even when predictions are compared relative to the same initial bed, the characteristics of the $\mathrm{MSESS}_{\mathrm{ini}}$ and its decomposition could lead to a preference for a prediction that is not consistent with the evaluator's judgement.

In summary, for $0 < \rho_{p'o'} < 1$, the amplitude error β' is minimized and, unless compensated by systematic bias γ', the $\mathrm{MSESS}_{\mathrm{ini}}$ is maximized for $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'}$, thus for predictions that underestimate the variance of the bed changes. Note that, similarly, the MSE can be minimized through an underprediction of the variance of bed levels[3]. Clearly, these findings have implications for (automated) calibration as well as validation procedures that minimize MSE or maximize[4] $\mathrm{MSESS}_{\mathrm{ini}}$.

2.3 Illustration for the real-life case of Bornrif

In this section, the conventional validation method, discussed in Sect. 2.2, is applied to 15 years of Delft3D (Lesser et al., 2004) morphodynamic computations for the Bornrif, a dynamic attached bar at the North-Western edge of the Wadden Sea barrier island of Ameland, the Netherlands. We specifically explore the correspondence between predictive skill as perceived by the $\mathrm{MSESS}_{\mathrm{ini}}$ and its decomposition on the one hand, and by visual validation on the other hand. Here, visual validation is considered as the diagnosis of prediction quality by visual inspection, which is a powerful yet qualitative and subjective validation method. First, Sect. 2.3.1 briefly describes the available observations and model set-up. Next, Sects. 2.3.2 and 2.3.3 evaluate the model results by visually inspecting the predicted and observed morphology and morphological change and by applying the conventional error statistics, respectively. In Sect. 2.3.4, the effect of the validation approach on the perception of model skill is further examined. Finally, the effect of spatial scales on the skill trend, as perceived by the $\mathrm{MSESS}_{\mathrm{ini}}$, is examined in Sect. 2.3.5.

2.3.1 Bornrif model and validation set-up

We have gratefully made use of available morphodynamic simulations from 1993 to 2008 (Achete et al., 2011), which were performed with the specific goal to hindcast the spit evolution at the Bornrif area and to project the findings to the Sand Engine pilot project at the Delfland coast (Stive et al., 2013). Only sediment transport due to waves and wave-induced currents was considered. To this end, a set of 12 wave conditions, representing the yearly-averaged climate, was applied throughout the simulation. While the horizontal tide and the dynamics of the adjacent ebb

[3] i.e. for $\sigma_p/\sigma_o = \rho_{po}$ with $0 < \rho_{po} < 1$, or, equivalently, for predictions that underestimate the variance of the bed changes, i.e. for $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'}$ with $0 < \rho_{p'o'} < 1$ (addendum to Bosboom et al., 2014).

[4] The word "maximize" was erroneously omitted in Bosboom et al. (2014), which is corrected here.

tidal delta were neglected, the vertical tide was taken into account. The morphodynamic evolution was computed on a grid with a resolution of 50 × 50 m^2 near the spit and 100 × 50 m^2 closer to the model boundaries. The initial bed for the simulations (Fig. 2.2) was prepared from the Vaklodingen data set (Wiegman et al., 2005).

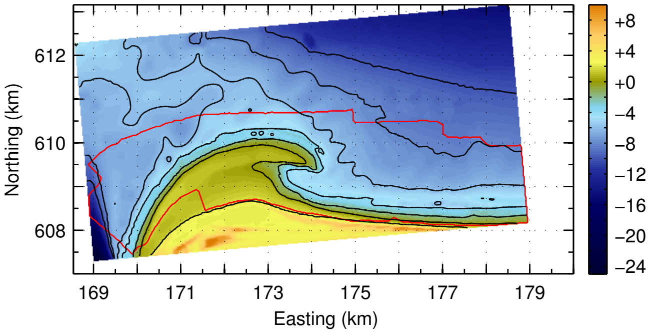


Figure 2.2: Initial bathymetry for the Bornrif simulation (1993) with the red polygon indicating the analysis region.

For the present validation, yearly bathymetric data up to depths of about 16m (JARKUS data; Minneboo, 1995) are available, interpolated to a 20 × 20 m^2 grid. The JARKUS measurements are more frequent than the Vaklodingen, but extend to smaller water depths. The measurements for 1994 were excluded from the analysis because of a significant gap in the data in the considered domain. In order to retain all observed scales, the comparison between the observed and computed fields is performed on the 20 × 20 m^2 grid that the JARKUS data were presented on. To that end, the computations were interpolated onto the observational grid. The red polygon in Fig. 2.2 delineates the overlap of the computational domain and the yearly observations during the entire period and defines the analysis region for which the various statistics are computed (see Sect. 2.3.3).

2.3.2 Visual validation

The bathymetries and the yearly and cumulative sedimentation/erosion fields within the bounding polygon are shown in Figs. 2.3 to 2.5, respectively. Visual validation of bathymetries shows that the computed general migration direction, the progressive attachment of the spit to the mainland and the subsequent infilling of the bay qualitatively correspond to the observations (Fig. 2.3). From about 1998, migrating sand bars are observed at water depths larger than 5 m to the east of the Bornrif, which are not reproduced by the model. The observations further differ from the computations in that a stronger and faster development and flattening of the overall shape takes place in reality. The rate between eastward and southward propagation is smaller in the computations leading to a shorter spit and a faster

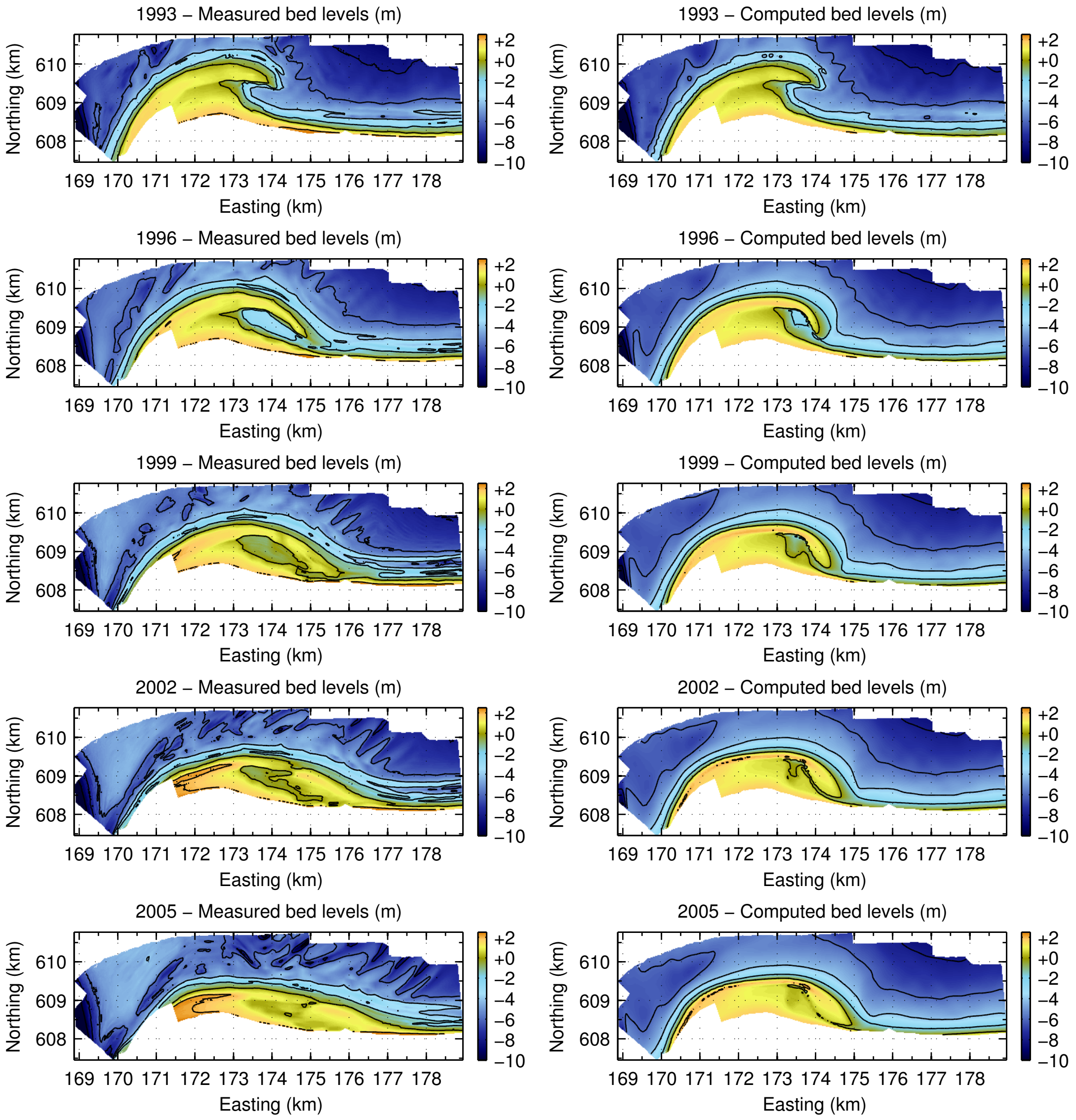

1993 – Measured bed levels (m)
1993 – Computed bed levels (m)
1996 – Measured bed levels (m)
1996 – Computed bed levels (m)
1999 – Measured bed levels (m)
1999 – Computed bed levels (m)
2002 – Measured bed levels (m)
2002 – Computed bed levels (m)
2005 – Measured bed levels (m)
2005 – Computed bed levels (m)
Northing (km)
610
609
608
Easting (km)
169 170 171 172 173 174 175 176 177 178
+2
+0
−2
−4
−6
−8
−10

Figure 2.3: Measured (*left*) and computed (*right*) Bornrif bathymetries for the years 1993, 1996, 1999, 2002 and 2005 for the analysis region.

land attachment (i.e. at a smaller alongshore distance) and a smaller bay. The visual comparison of computed and observed bathymetries suggests a decreasing correspondence in time.

The observed yearly sedimentation/erosion fields (Fig. 2.4) are very different from the computed fields in that they show a strong, small-scale morphological variability, not reproduced by the model, in the larger part of the domain. The

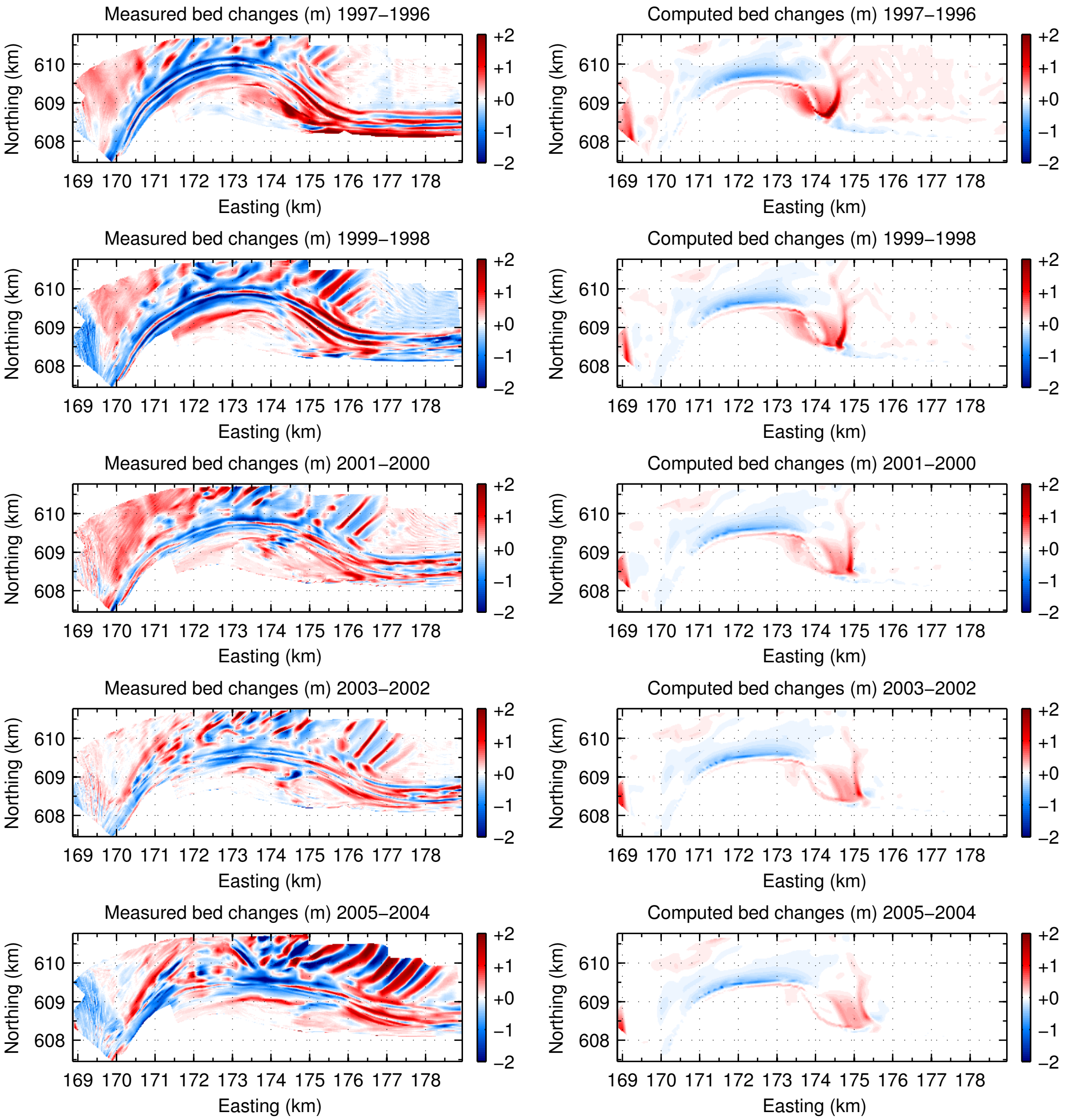


Figure 2.4: Measured (*left*) and computed (*right*) Bornrif yearly bed changes for several years.

strength of this variability changes significantly from year to year. From 1998, the sand bars are clearly visible, particularly at larger water depths to the east of the Bornrif. In the inlet channel, alternating sedimentation and erosion is observed, whereas the computations show consistent sedimentation. The visual agreement between measured and computed yearly bed changes is limited in all years. The magnitude of the changes is best represented at the start of the computations and deteriorates with time, as the computed yearly changes strongly reduce towards

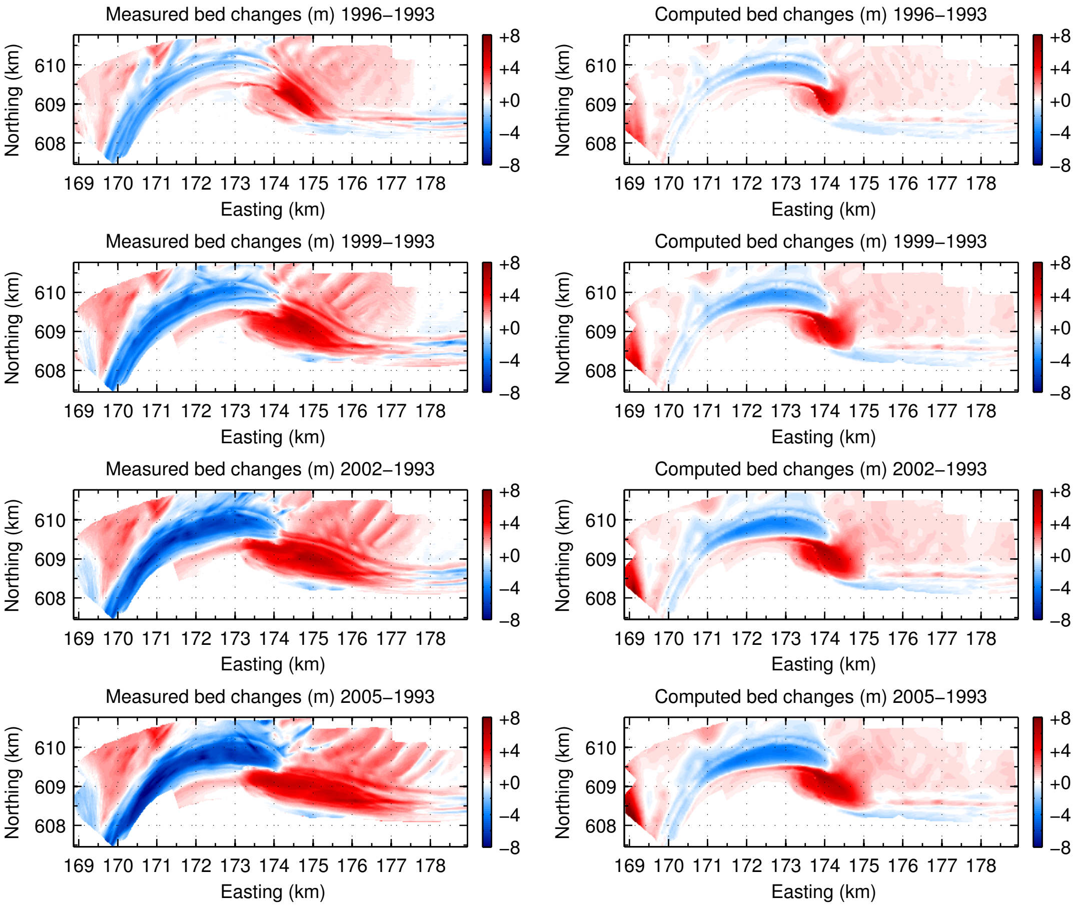


Figure 2.5: Measured (*left*) and computed (*right*) Bornrif cumulative bed changes for 1996, 1999, 2002 and 2005 with respect to the initial bed of 1993.

the end of the simulation.

The cumulative bed changes (Fig. 2.5) show that the model qualitatively reproduces the main nearshore feature of large-scale erosion and sedimentation in the western and eastern parts of the domain, respectively. The spatial extent and the overall magnitude of the cumulative changes, however, are significantly larger in the observations, and increasingly so in time. Another marked difference between observations and computations is that the observed pattern shifts eastward with time, whereas the computed pattern remains more localized. The computations further show net sedimentation in the inlet channel that is not found in reality. The migrating sand bars are best recognized from the yearly changes, but are also visible in the observed fields of cumulative change, where they are evident as a smaller-scale variation to the larger-scale trend.

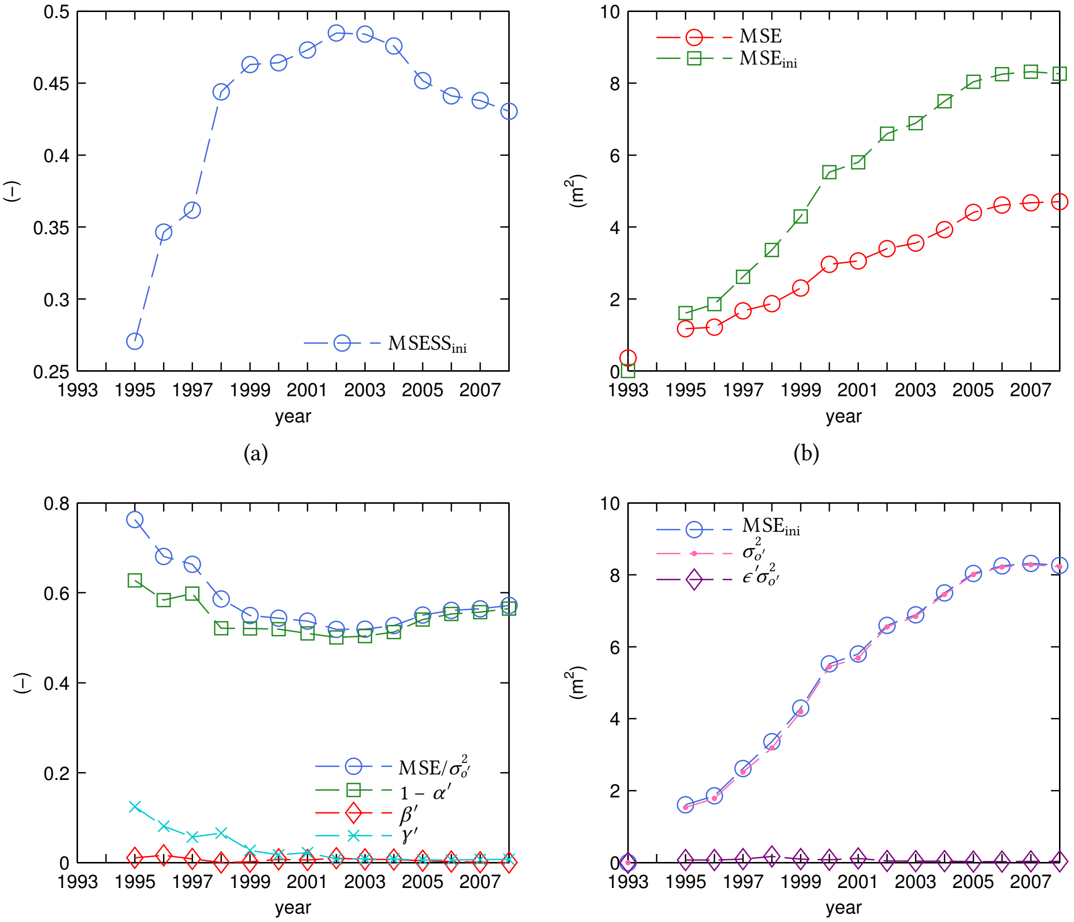
By definition, the point-wise error $(p - o) = (p' - o')$. Nonetheless, while it was easily concluded that the quality of the bathymetric fields (Fig. 2.3) deteriorates with time, it is much harder to visually judge the quality of the cumulative sedimentation/erosion fields over time (Fig. 2.5). On the one hand, the underestimation of the overall magnitude of the bed changes can be seen to rapidly increase in time, at least until 2002. On the other hand, the centres of cumulative erosion, which attract immediate attention, seem to be located closer together in for instance 2002 than in 1996. This ambiguity (and its absence for bed levels) is further explored in Sect. 2.3.4 by comparison with the conventional error statistics that are discussed in the next section.

2.3.3 Conventional error statistics

The skill score $\mathrm{MSESS}_{\mathrm{ini}}$ according to Eq. 2.11 is the lowest at the beginning of the simulation and gradually increases over time from the start of the simulation until 2002, after which the skill slightly decreases again (Fig. 2.6a). According to Table 2.1, the score qualifies as “good” for all years. Based on $\mathrm{MSESS}_{\mathrm{ini}}$, we would conclude that the quality of the predictions increases with time, at least for the main part of the simulation until 2002. In contrast, the accuracy of the modelled bed levels, or equivalently, of the sedimentation/erosion fields decreases with simulation time, evident from the increase in MSE (Fig. 2.6b)[5]. That nonetheless the skill, viz. the relative accuracy, increases with time is due to $\mathrm{MSE}_{\mathrm{ini}}$, the MSE of the reference prediction, increasing with time and, until 2002, at a faster rate than the MSE of the predictions (Fig. 2.6b). With $\mathrm{MSE}_{\mathrm{ini}} = \langle \Delta z_o^2 \rangle$, its behaviour is governed by the increase of the mean-squared cumulative observed bed changes as a result of the natural development away from the initial situation.

Figures 2.6a and 2.6b exemplify that, for a trend, the accuracy required for a certain level of skill decreases further into the simulation (Sect. 2.2.3). In order to better value $\mathrm{MSESS}_{\mathrm{ini}}$ and its temporal variation, a detailed analysis is needed of the terms that contribute to the absolute and relative accuracy. The decomposed error terms as defined through Eqs. 2.6 and 2.7 and Eqs. 2.8 and 2.9, with $p' = \Delta z_p$ and $o' = \Delta z_o$, are shown in Fig. 2.6c and Fig. 2.6d, respectively. The MSE normalized with the variance of the observed anomalies, shown in Fig. 2.6c, is dominated by the phase error $1 - \alpha'$ of the anomalies. The normalized sediment budget error γ' decreases with time and only plays a role in the first half of the simulation, while the amplitude error β' is negligible throughout the simulation. Figure 2.6d illuminates that the bias part $\epsilon' \sigma_{o'}^2$ of $\mathrm{MSE}_{\mathrm{ini}}$ is negligible ($\epsilon' \ll 1$), such that $\mathrm{MSE}_{\mathrm{ini}} \approx \sigma_{o'}^2$. The skill score (Eq. 2.10) is thus given by $\mathrm{MSESS}_{\mathrm{ini}} \approx 1 - \mathrm{MSE}/\sigma_{o'}^2 \approx \alpha' - \gamma'$ and from, say, 1999, $\mathrm{MSESS}_{\mathrm{ini}} \approx \alpha'$. Thus, the decrease of both the phase error $1 - \alpha'$ and the sediment budget error γ' contributes to the increase

[5] Note that the MSE is not exactly zero for the simulation start time due to the Delft3D algorithm applied to interpolate the 1993 observed bathymetry to the water-depth points of the staggered computational grid.



(a)

(b)

(c)

(d)

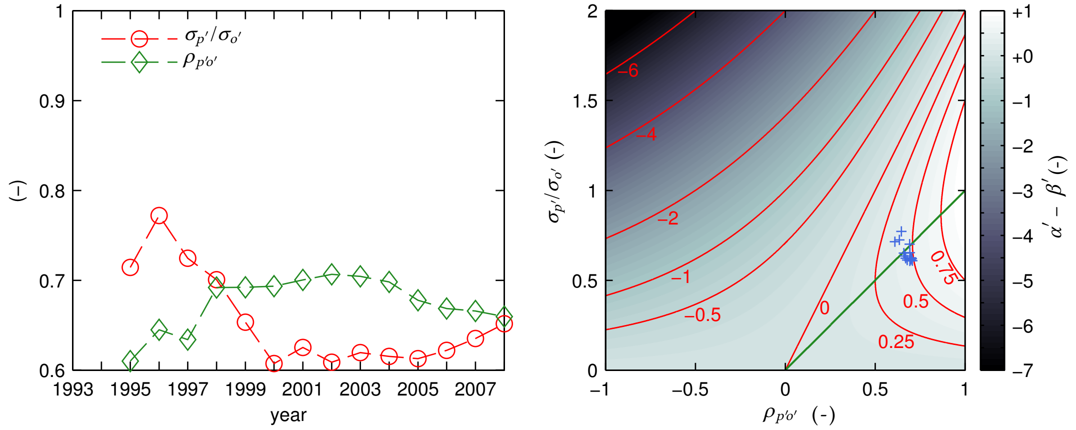
Figure 2.6: Model performance for Bornrif: **(a)** MSE skill score with the zero change model as the reference, MSESS$_{ini}$, **(b)** MSE of the computations and MSE$_{ini}$ of the initial bed (zero change reference model), **(c)** MSE normalized with the variance of the cumulative observed bed changes and its decomposition, Eqs. 2.6 and 2.7 and **(d)** MSE$_{ini}$ and its decomposition, Eqs. 2.8 and 2.9.

in skill until 2002, the year that exhibits most skill as well as the smallest phase error. From 2002–2003 onwards, the phase error increases and, consequently, the skill decreases. Below, we further explain these findings.

The sediment budget error γ' normalizes an absolute map-mean error MSE$_{\text{bias}}$ = $(\overline{p} - \overline{o})^2 = (\overline{p'} - \overline{o'})^2$ with the variance of the cumulative observed bed changes $\sigma^2_{o'}$ (Eq. 2.7c). Analysis showed that the rapid decrease of γ' until 2000 is mainly due to the strong increase of $\sigma^2_{o'}$ over time rather than through variation of MSE$_{\text{bias}}$.

The negligible amplitude error β' (Eq. 2.7b) is the direct result of $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$ being relatively close together in value (Fig. 2.7a) and is not to be interpreted as an indicator that the correct volumes of sand are moved; pair-wise comparison of the observed and computed fields of cumulative change (Fig. 2.5) suggests a consistent and over time increasing underprediction of the magnitude of the cumulative bed changes, and, thus, of the volumes of sand moved, at least in the first half of the simulation (see also Sect. 2.2.4). This is confirmed by the behaviour of the ratio $\sigma_{p'}/\sigma_{o'}$ between the standard deviations of computed and measured cumulative bed changes, which has values consistently smaller than 1 and as low as about 0.6 from 2000 onwards (Fig. 2.7a).

The effect on the skill score is visualized in Fig. 2.7b, which shows the behaviour of $\text{MSESS}_{\text{ini}} = \alpha' - \beta'$ (Eq. 2.10 assuming $\gamma' = \epsilon' = 0$) as a function of $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$. As expected, the values for the Bornrif simulation can be seen to lie close to the green diagonal ($\rho_{p'o'} = \sigma_{p'}/\sigma_{o'}$) along which β' is minimized. Consequently, for the Bornrif, a much smaller underestimation of the variance of the cumulative bed changes would, counter-intuitively, have raised MSE values and lowered the diagnosed skill levels, as in the case of East Pole Sand (Sect. 2.2.4).



(a)

(b)

Figure 2.7: Skill levels benefit from underestimation of bed changes: **(a)** correlation $\rho_{p'o'}$ and ratio of the standard deviations $\sigma_{p'}/\sigma_{o'}$ of the predicted and observed cumulative bed changes for the Bornrif simulation, **(b)** skill score $\text{MSESS}_{\text{ini}} = \alpha' - \beta'$ (assuming $\gamma' = \epsilon' = 0$) as a function of $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$ with the Bornrif values for all years indicated with "+". Along the green diagonal ($\rho_{p'o'} = \sigma_{p'}/\sigma_{o'}$), the amplitude error is minimized and, in the absence of map-mean errors, the skill maximized at $\text{MSESS}_{\text{ini}} \approx \alpha'$.

With β' negligible and γ' vanishing after the first years of the simulation, the skill score $\text{MSESS}_{\text{ini}}$ peaks simultaneously with the phase association α' and the maximum value of $\text{MSESS}_{\text{ini}}$, in 2002, is fully determined by α' (Figs. 2.6a and 2.6c). In Sect. 2.2.2, we interpreted α' as the structural similarity between predicted and observed cumulative sedimentation/erosion patterns. Since it is invariant to map-mean error and changes in scale of observations and predictions (in other words: the mean and variance of observed and predicted bed changes are irrelevant), α'

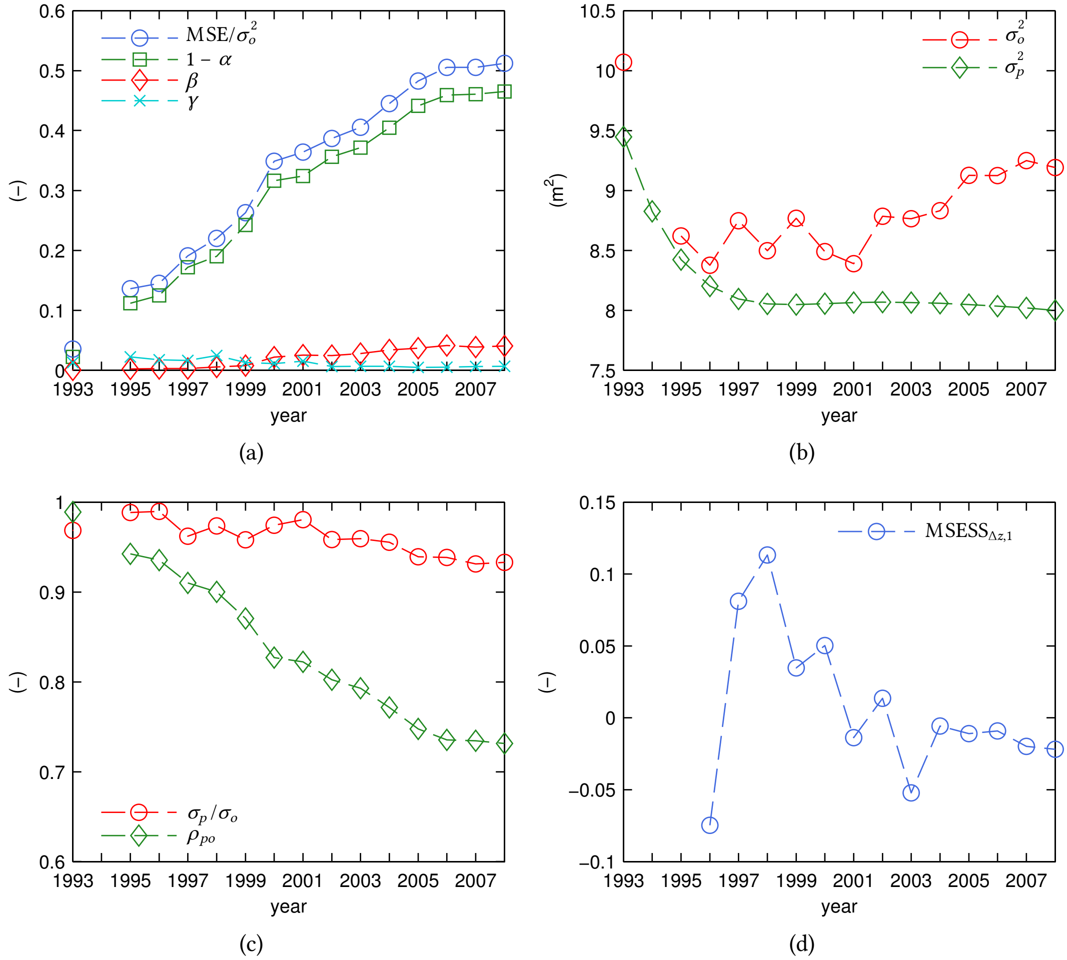
does not provide information on the accuracy of predictions (Willmott, 1982).

In summary, it is inherent to the use of the initial bed as the reference that while the morphology progressively develops away from the initial bed, larger absolute errors (MSE, $\mathrm{MSE_{bias}}$) are allowed in order to obtain a certain level of skill. Further, for the Bornrif simulation, the skill levels benefit from the consistent underestimation of the magnitude of the bed changes ($\sigma_{p'}/\sigma_{o'} < 1$). In fact, the underestimation is largest in 2002, the year for which maximum skill is reported. This undesirable behaviour of $\mathrm{MSESS_{ini}}$ is inherited from the use of the MSE as the accuracy measure (cf. Sect. 2.2.4). The skill maximum is due only to the greatest similarity, in 2002, in the structure of the sedimentation/erosion patterns (as measured by $\rho_{p'o'}$ or α').

2.3.4 Visual validation versus error statistics

Sections 2.3.2 and 2.3.3 illustrated that prediction quality, the degree of correspondence between predictions and observations (Murphy, 1993), is a multidimensional concept. Logically, as follows from Eq. 2.10, $\mathrm{MSESS_{ini}}$ and its components describe aspects of prediction quality related to the cumulative sedimentation/erosion fields from the start of a simulation. While visually judging fields of cumulative change, we tend to compare the structure as well as the magnitude of the fluctuating parts of pairs of observations and predictions (Sect. 2.3.2). A small bias, as in Fig. 2.5, will most likely go unnoticed. Our impression, from Fig. 2.5, of the structure and magnitude of the anomalies over time qualitatively corresponds to the behaviour of $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$ (Fig. 2.7a), respectively. The opposite behaviour of $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$ explains the ambiguity that was found in visually judging, based on Fig. 2.5, whether the predictions in 1996 or 2002 are of higher quality. On the contrary, the development of $\mathrm{MSESS_{ini}}$ over the course of the simulation was seen to merely report the correlation $\rho_{p'o'}$ between cumulative sedimentation/erosion fields (Fig. 2.6c), such that the 2002 predictions are diagnosed with maximum skill (Fig. 2.6a). A morphologist, however, asked to visual judge the fields of cumulative change, will probably only reach a similar conclusion when turning a blind eye to the differences in scale, both between observations and predictions at a particular time and between pairs of observations and predictions at different times.

Prediction quality, as perceived by pair-wise visual comparison of bed levels rather than cumulative change, was unambiguously found to deteriorate over time (Sect. 2.3.2). Clearly, even though $\mathrm{MSE} = \langle (z_p - z_o)^2 \rangle = \langle (\Delta z_p - \Delta z_o)^2 \rangle$, other aspects of prediction quality are highlighted when visually judging the closeness of bed levels instead of cumulative sedimentation/erosion fields. This can be explained by considering the Murphy–Epstein decomposition of MSE in terms of the bed levels (Eq. 2.14 and Fig. 2.8a), as opposed to of the anomalies (Eq. 2.6 and Fig. 2.6c). Although the variance of the observations σ_o^2 varies in time (Fig. 2.8b),



(a)

(b)

(c)

(d)

Figure 2.8: Comparison of overall statistics for measured and computed Bornrif bathymetries: **(a)** MSE normalized with the observation variance and its decomposition, **(b)** variance of observed and computed bed levels, **(c)** correlation and ratio of standard deviations of the measured and predicted bed levels and **(d)** MSE skill score for yearly bed changes with a zero change reference model (MSESS$_{\Delta z,1}$). Note that, similar as for the MSE[5], the 1993 measured and computed parameters differ slightly.

it is relatively constant as compared to $\sigma_{o'}^2$ (Fig. 2.6d). Hence, where the MSE normalized with $\sigma_{o'}^2$ behaves quite differently from the MSE itself, the MSE normalized with σ_o^2 increases in time as the MSE does. From Fig. 2.8a, MSE/σ_o^2 can be seen to be dominated by the phase error $1-\alpha$, which increases with time as a result of the decreasing correlation ρ_{po} between predicted and observed bed levels (Fig. 2.8c). Analogously, the most obvious finding from the visual validation of bed levels (Fig. 2.3) was the decreasing overall agreement in structural similarity between the measured and predicted bathymetric fields. The slight increase in amplitude error β is governed by the fact that ρ_{po} decreases faster with time than σ_p/σ_o (Fig. 2.8c). Note further that, analogously to β', β would have been larger, if only slightly, for $\sigma_p/\sigma_o = 1$.

The normalized metrics for the bed levels, ρ_{po} and σ_p/σ_o, provide information not contained in the anomalies. For instance, from Fig. 2.8b, it is apparent that the computational variance develops towards a constant, too low level at which the larger-scale modelled bathymetry appears to be in equilibrium with the applied representative yearly-averaged wave climate. Further, without taking possible compensation due to systematic bias into account, $\rho_{po} < \sigma_p/\sigma_o$ indicates that at deeper water the predicted depths are overestimated (and at smaller depths underestimated), see Appendix 2.A. A regression demonstrated that this is most likely the result of the large extent of sedimentation at deeper water that is not mimicked by the model (Fig. 2.3).

In conclusion, $\text{MSESS}_{\text{ini}}$ by itself sheds a limited light on the model performance for the Bornrif; it merely reports the development of the correlation $\rho_{p'o'}$ between cumulative sedimentation/erosion fields. A morphologist, asked to visually evaluate the time evolution of model performance on the basis of Figs. 2.3 and 2.5, would most likely report his impression of the degree of overall correspondence between the fields, the relative role of map-mean error and the extent to which the magnitudes and structure of the fields of cumulative change and bed levels are reproduced. These subjective notions can be quantified by e.g. MSE, MSE_{bias}, $\sigma_{p'}/\sigma_{o'}$, $\rho_{p'o'}$, σ_p/σ_o and ρ_{po} respectively.

2.3.5 The effect of various spatial scales

The various statistics, discussed in Sect. 2.3.4, inevitably combine information across a range of spatial scales. Hence, it is nontrivial to relate ρ_{po} and σ_p/σ_o or $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$ to particular features of interest in the morphology or the fields of cumulative change, respectively. The range over which spatial scales are lumped together is especially wide for the normalized bed level metrics, ρ_{po} and σ_p/σ_o, in which scales up to the size of the model domain play a role (cf. Sect. 2.3.4). By implication, the values of ρ_{po} or σ_p/σ_o are sensitive to the inclusion of morphologically inactive regions, which is not the case for $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$, and are arguably dominated by the larger scales.

Upon visual inspection, it was concluded that the simulations capture little of the year-to-year variability, while the larger-scale fields of cumulative change are reasonably well predicted (Sect. 2.3.2). This suggests that the relative contribution of the smaller scales to $\rho_{p'o'}$, $\sigma_{p'}/\sigma_{o'}$ and $\text{MSESS}_{\text{ini}}$ decreases during the simulation.

The skill at smaller spatial scales can be quantified by taking a slightly different approach to skill, which considers the bathymetric *change* rather than the morphology itself. Skill can now be defined as the relative accuracy of bed changes rather than bed levels, using a reference of zero change and considering bed changes in a one-year period. Denoting the yearly predicted and measured bed changes with $\Delta z_{p,1}$ and $\Delta z_{o,1}$, respectively, we now have $p = \Delta z_{p,1}$ and $o = \Delta z_{o,1}$ in Eq. 2.3 and $r = 0$ in Eq. 2.5. Upon substitution, Eq. 2.2 yields:

$$\mathrm{MSESS}_{\Delta z,1} = 1 - \frac{\left\langle \left(\Delta z_{p,1} - \Delta z_{o,1}\right)^2 \right\rangle}{\left\langle \Delta z_{o,1}^2 \right\rangle}. \tag{2.12}$$

Note that if the period of bed changes were taken as the simulation duration up to the evaluation time, we obtain $\mathrm{MSESS}_{\mathrm{ini}}$ (Eq. 2.11).

For all years of the Bornrif simulation, the relative accuracy of yearly change, $\mathrm{MSESS}_{\Delta z,1}$, is low or negative (Fig. 2.8d) and tends to decrease further into the simulation. Note that the relatively low value for 1996 is the result of the rather small observed morphological change in 1995–1996 (Fig. 2.6d). Since Eq. 2.12 does not consider any cumulative effect on timescales larger than one year, the cancellation of errors over the course of multiple years (as can be expected specifically for the smaller spatial scales) is not taken into account.

Based on the above, we hypothesize that the relatively low values of $\mathrm{MSESS}_{\mathrm{ini}}$ at the beginning of the Bornrif simulation (Fig. 2.6a) are mainly due to unskilful smaller spatial scales. When, over time, the relative contribution of these smaller scales to the cumulative change decreases, the larger scales are allowed a greater opportunity to become correlated to the predictions, until at some point in the simulation, the main part of the skill is attributable to the more skilful, persistent large-scale trend. Hence, further into the simulation, on average higher skill values are found. The same phenomenon may also, at least partly, explain the period of negative to low skill that is referred to as spin-up time and often found at the beginning of long-yearly morphodynamic simulations (Dam et al., 2013). An increase in skill, for longer prediction horizons, is then to be interpreted as the emerging of the more skilful larger scales. Clearly, the above demonstrates the need for validation methods that distinguish between various spatial scales.

2.4 Summary and discussion

The use of $\mathrm{MSESS}_{\mathrm{ini}}$ (Eq. 2.11) as (the main) indicator of morphodynamic model performance has implications for the perception of model skill. We summarize and discuss these implications in this section. First, Sect. 2.4.1 focuses on the effect of the choice of the zero change reference model. Second, Sect. 2.4.2 summarizes the aspects of model performance captured by $\mathrm{MSESS}_{\mathrm{ini}}$ as well as by visual validation.

2.4.1 The zero point at the scale of skill

The $\mathrm{MSESS}_{\mathrm{ini}}$ is frequently used to compare morphodynamic model performance across different prediction situations. We have demonstrated however, that the validity of the ranking based on $\mathrm{MSESS_{ini}}$ (Table 2.1) is limited and that absolute values of skill levels for different geographical locations, time periods or forcing conditions should not be compared. For the $\mathrm{MSESS_{ini}}$ to create a level playing field, the cumulative observed bed changes from the initial bed must adequately reflect the intrinsic difficulty levels across situations with a different morphological development (for instance trend, cyclic/seasonal, episodic or combinations thereof). Synthesized examples (Sect. 2.2.3) showed that this assumption cannot be expected to hold.

In connection with the above, it was argued that $\mathrm{MSESS_{ini}}$ may also misreport the temporal evolution of model skill. For inter-seasonal modelling of seasonal systems, the normalization with the mean-squared cumulative bed changes may result in an artificial seasonal variation of the accuracy of the initial bed and hence of the reported model skill (Sect. 2.2.3). More in general, when predicting cyclic morphodynamics, any single-state reference, whether a longer-term average or an arbitrary moment's actual bathymetry, unavoidably leads to a zero level on the scale of skill that fluctuates with the observed deviation from the reference.

For prediction situations that include a trend, the use of the zero change reference model means that, in time, the minimal level of acceptable performance is lowered at a rate determined by the cumulative observed bed changes (Sect. 2.2.3). If the accuracy of the reference model decreases in time at a faster rate than the accuracy of the predictions, the $\mathrm{MSESS_{ini}}$ may even increase with time, while the agreement between modelled and observed bathymetry strongly decreases, as was seen for the Bornrif (Sect. 2.3.3). It is debatable whether the zero change reference model sets an ambitious enough quality standard, especially for longer prediction horizons. For instance, the 2008 Bornrif prediction obtains positive skill if it outperforms the prediction "2008 is like 1993", 1993 being the start of the simulation (Sect. 2.3.3). This reference prediction, however, is not very likely in the eyes of a morphologist, who expects the Bornrif to gradually diffuse eastward.

A slightly different normalization is applied by Ruessink and Kuriyama (2008), who normalize with the *expected value* of the mean-squared difference between two bathymetric profiles with a sampling interval equal to the time elapsed from the start of the simulation. Although in this way the accuracy of the zero change reference is determined in an averaged sense, the magnitude of the denominator remains dependent on the cumulative morphological development.

Alternatives to the model of zero change, valid across different morphological systems, are nontrivial. For inter-seasonal modelling of seasonal systems, a persistence model could be adequate as long as the observations from the same season are assumed to persist (as opposed to assuming that the initial bed persists). If for the example of the summer–winter cycle in Sect. 2.2.3, the initial or last observed state from the same season were used, this would have eliminated the artificial seasonal fluctuation of the accuracy of the reference and subjected the summer and winter profiles to an equal test. Naturally, for a trend, a more appropriate naive model would be some estimate of the trend, producing more accurate reference predictions than the zero change model. One of the rare examples in morphodynamic modelling is due to Davidson et al. (2010) who make use of a linear trend prediction as the benchmark for coastline modelling. Unfortunately, for area models the quantification of a naive trend prediction is far from trivial.

In conclusion, a comparative evaluation based on skill scores, however defined, is unlikely to have general validity. Instead of through an absolute ranking of predictions, skill levels should thus be valued on a case-by-case basis. In doing so, when reporting the temporal variation of $\mathrm{MSESS_{ini}}$, we recommend that at the very least also values of MSE are reported, such that a broader view on model performance can be obtained than by using $\mathrm{MSESS_{ini}}$ alone.

2.4.2 Multiple dimensions to prediction quality

Using the evaluation of the Bornrif model performance as an example, multiple aspects of prediction quality were identified, viz. the extent to which the magnitudes and structure of the fields of cumulative change and bed levels are reproduced, the degree of overall correspondence between the fields and the relative role of map-mean error (Sect. 2.3.2). These notions can be quantified by e.g. $\sigma_{p'}/\sigma_{o'}$, $\rho_{p'o'}$, σ_p/σ_o, ρ_{po}, MSE and $\mathrm{MSE_{bias}}$, respectively (Sects. 2.3.3 and 2.3.4). Summary metrics, such as the MSE and the $\mathrm{MSESS_{ini}}$, were seen to provide an implicit weighting of systematic bias terms as well ρ_{po} and σ_p/σ_o and $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$, respectively. Unfortunately, in doing so, MSE and $\mathrm{MSESS_{ini}}$ tend to reward the underprediction of the variance of bed levels and bed changes, respectively, as shown in Sect. 2.2.4.

This tendency of the mean-squared-error measure of accuracy, in combination with the model of zero change, to favour predictions that underestimate the variance of the cumulative bed changes, was easiest appreciated in the absence of systematic bias and sediment import or export ($\gamma' = \epsilon' = 0$). Then, $1 - \mathrm{MSESS_{ini}}$ differs from MSE by a factor $1/\sigma_o'^2$ and is fully determined by the correlation $\rho_{p'o'}$ and the ratio of the standard deviations $\sigma_{p'}/\sigma_{o'}$ of the predicted and measured bed changes. It was found that for the same map-mean errors and suboptimal $\rho_{p'o'}$ ($0 < \rho_{p'o'} < 1$), the skill $\mathrm{MSESS_{ini}}$ is maximized for $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'}$, hence for too small overall bed changes (Sect. 2.2.4 and Fig. 2.7b). For a real-life case, taken from literature, this was shown to have resulted in the ranking of predictions based on $\mathrm{MSESS_{ini}}$ being inconsistent with expert judgement (Sect. 2.2.4). Similarly, since for the Bornrif simulation $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$ are close together in value (Fig. 2.7a), the skill levels are dominated by $\rho_{p'o'}$. As a result, the development of $\mathrm{MSESS_{ini}}$ in time was seen to merely report the correlation $\rho_{p'o'}$ between cumulative sedimentation/erosion fields (Sect. 2.3.3) and the year with the largest underestimation of the variance of cumulative change could be diagnosed with maximum skill.

Clearly, this finding has implications for (automated) calibration procedures that minimize $\mathrm{MSESS_{ini}}$; for positive, suboptimal correlation, reduction of the overall sizes of bed changes by, for instance, choosing an unrealistic transport parameter is an effective, though undesirable method to obtain higher values of $\text{MSESS}_{\text{ini}}$.

In morphodynamic model validation, the $\text{MSESS}_{\text{ini}}$ is sometimes supplemented with its Murphy–Epstein decomposition (Eq. 2.10). Although this may provide some of the required extra information, a few warnings are warranted here. First, the phase and amplitude errors according to the Murphy–Epstein decomposition, $1-\alpha'$ and β', respectively, are not necessarily in line with the morphologists' intuitive definition. The phase association α' (Eq. 2.7a) is best explained as a measure of the structural similarity between the sedimentation/erosion fields, indicating to what extent not only locations but also shapes and relative magnitudes of the sedimentation/erosion features are correct (but note that α' does not distinguish between positive and negative correlations). Further, when neglecting systematic bias, $\sigma_{p'}/\sigma_{o'}$ rather than β' (Eq. 2.7b) would be the more appropriate overall indicator of agreement between the predicted and observed sizes of bed changes and, therefore, cumulative volumes of sand moved. Finally, the interpretation of the sediment budget error γ' (Eq. 2.7c) is also nontrivial, since it normalizes an absolute sediment budget error with the variance of the cumulative observed bed changes. This normalization, and the related complications for the interpretation of γ', are inherited from the zero change reference model.

None of the above mentioned measures facilitates a distinction between the multiple scales at which features of interest appear in bed levels and fields of cumulative change. As a consequence, they do not provide guidance as to which scales in the output can be considered of sufficient quality. Furthermore, their temporal variation may carry the signature of a combination of small-scale variability and larger-scale trends. For instance, negative or low values of $\text{MSESS}_{\text{ini}}$ at the beginning of a simulation may be attributable to inadequately represented small-scale variability, whereas larger values further into the simulation could be due to larger-scale trends (Sect. 2.3.5).

In summary, although frequently used as the main indicator of morphodynamic model skill, the use of $\text{MSESS}_{\text{ini}}$ (or any other measure of quality) is not sufficient to describe prediction quality in its full dimensionality. In order to capture the various aspects of model performance contained in the fields of bed levels and cumulative and yearly sedimentation/erosion, multiple accuracy/skill measures must be reported (Sects. 2.3.3 and 2.3.4). In doing so, it is crucial, yet nontrivial, to fully appreciate which aspect(s) of model quality is (are) exactly captured in a particular score. A method that allows any metric to selectively address multiple spatial scales could further broaden our view on model performance (see Bosboom and Reniers, 2014a, i.e. Ch. 6). Finally, the tendency of MSE and $\text{MSESS}_{\text{ini}}$ to reward the underprediction of the variance of bed levels and bed changes, respectively, calls for the development of alternative summary metrics (e.g. Taylor, 2001; Koh et al., 2012; Bosboom and Reniers, 2014a,b, i.e. Chs. 4 and 6).

2.5 Conclusions and future work

As demonstrated with synthetic examples, examples from literature and a long-yearly Delft3D model simulation, the mean-squared-error skill score relative to a prediction of zero change may produce a relative ranking of predictions that does not match the intuitive judgement of experts. This is true for the comparison of skill across different prediction situations, e.g. different forcing conditions or internal dynamics, as well for the temporal variation of skill within a simulation. Two main causes of unexpected skill are identified. First, the zero change reference model assumes that the conditions at the start of the simulations persist in time, such that the minimal level of acceptable performance varies with the mean-squared observed cumulative change. The latter fails to reflect the relative difficulty of prediction situations with a different morphological development prior to the evaluation time (for instance trend, cyclic/seasonal, episodic or combinations thereof). Second, since the MSE is prone to reward predictions that underestimate variability, an underprediction of the variance of cumulative bed changes leads to a higher diagnosed skill.

On a case-by-case basis, a balanced appreciation of model performance requires that multiple accuracy and/or skill metrics are considered in concert. For instance, the temporal evolution of skill as diagnosed through the mean-squared-error skill score is best valued in combination with the MSE itself. In addition, we recommend the use of separate measures for map-mean error and magnitude and structure of the fluctuating parts, for both morphology and bed changes, which are more in line with the morphologists' intuitive definition than the decomposed error contributions according to the Murphy–Epstein decomposition.

Of course, the morphologist may sometimes still desire a single-number summary of the main aspects of model performance, especially if automated calibration routines are used. We are therefore exploring alternative summary metrics that, unlike grid-point based accuracy measures, such as the MSE, and its derived MSE skill score relative to the initial bed, penalize the underestimation of variability. For instance, experimental work is undertaken to formulate error metrics that take the spatial structure of 2D morphological fields into account (Bosboom and Reniers, 2014b, i.e. Ch. 4). Further, since model predictions are not necessarily of similar quality at different spatial scales, a method is being developed that allows any metric to selectively address multiple scales (Bosboom and Reniers, 2014a, i.e. Ch. 6). This scale-selective validation method for 2D morphological predictions provides information on model skill and similarity in amplitude and structure per spatial scale as well as aggregated over all scales.

Acknowledgements

We thank Fernanda Minikowski Achete for providing us with her Delft3D simulations of the Ameland Bornrif. Further, we thank the reviewers for their constructive comments, which resulted in an improved paper.

2.A Murphy–Epstein decomposition of MSE

Algebraic manipulation of the MSE, Eq. 2.3, leads to (Murphy, 1988):

$$\text{MSE} = \sigma_p^2 + \sigma_o^2 - 2\sigma_p\sigma_o\rho_{po} + \left(\overline{p} - \overline{o}\right)^2 \tag{2.13}$$

where $\overline{p}$ and $\overline{o}$ are the weighted map means and σ_p and σ_o the weighted standard deviations of the predictions p and the observations o, respectively, and ρ_{po} is the weighted Pearson product-moment correlation between the predictions and the observations. The latter is given by $\rho_{po} = \sigma_{po}/(\sigma_p\sigma_o)$, with σ_{po} denoting the weighted covariance between p and o, and reflects the overall strength and direction of the linear correspondence between pairs of computations and observations; a deviation from −1 or 1 implies scatter around the best linear fit. We can rearrange the terms in Eq. 2.13 to arrive at (Murphy and Epstein, 1989):

$$\text{MSE} = \sigma_o^2(1 - \alpha + \beta + \gamma) \tag{2.14}$$

where

$$\alpha = \rho_{po}^2 \tag{2.15a}$$

$$\beta = \left(\rho_{po} - \frac{\sigma_p}{\sigma_o}\right)^2 \tag{2.15b}$$

$$\gamma = \frac{\left(\overline{p} - \overline{o}\right)^2}{\sigma_o^2}. \tag{2.15c}$$

Here, γ is a normalized map-mean error. The term β is the conditional bias, which is nonzero if the slope $b = \rho_{po}\sigma_o/\sigma_p$ of the regression line of the observations o, given the predictions p, deviates from 1. Given a positive correlation and unless compensated by systematic bias, $b > 1$ indicates that smaller values are overpredicted and larger values are underpredicted (and vice versa for $b < 1$). The term α is the coefficient of determination defined as the proportion of the variation in the values of o that can be linearly "explained" (in a statistical sense) by p, or vice versa (Taylor, 1990).

Since $\text{MSE} = \langle(p - o)^2\rangle = \langle(p' - o')^2\rangle$, Eqs. 2.13 to 2.15 are equally valid when p and o are replaced with p' and o', respectively.

3 The deceptive simplicity of the Brier skill score

This chapter is republished with minor changes only from J. Bosboom and A. Reniers (2018). The deceptive simplicity of the Brier skill score. In: Y.C. Kim (Ed.), *Handbook of Coastal and Ocean Engineering*, pp. 1639–1663, doi:10/c5tr.

The often used mean-squared-error skill score (MSESS), a.k.a. the Brier skill score (BSS) is based on the assumptions that the mean-squared error (MSE) is an appropriate measure of correspondence for morphological predictions and that the accuracy of the initial bed as the reference correctly reflects the inherent difficulty or ease of prediction situations. In Ch. 2 (Bosboom et al., 2014), it was demonstrated that unexpected skill may be reported due to a violation of either of these assumptions.

The goal of Ch. 3 is to further investigate and illustrate the behaviour of the BSS through numerous simple examples and examples from literature. Besides, we pay due attention to the evaluation of the treatment of measurement error in the skill scores and the skill rankings. In order to account for measurement error, adjusted MSESS formulations and skill classifications have been suggested by van Rijn et al. (2003) and Sutherland et al. (2004). Unfortunately, this has initiated an inconsistent use of skill definitions and rankings in subsequent literature. This chapter establishes the best method to take measurement error into account.

The highlights of this chapter are:

1. Simple examples demonstrate how the Murphy–Epstein decomposition of the MSESS must be interpreted.
2. Existing methods to correct for measurement error are shown to be inconsistent in either their skill formulation or their suggested classification scheme.
3. It is illustrated through various examples that the initial bed as the reference prediction does not succeed in making model performance comparable for common morphological prediction situations.
4. Hypothetical examples and examples from literature illustrate that if maximizing the MSESS is the objective, underpredicting the variability of bed changes is generally advantageous.
5. It is exemplified that the combination of larger, persistent and smaller, intermittent scales of cumulative change leads to an increase of the MSESS with time, while the skill on either of these scales is kept constant in time.

Abstract

The quality of morphodynamic predictions is often indicated by a skill score that weights the mean-squared error (MSE) of the prediction by that of the initial bed as the reference prediction. As simple as this Brier skill score (BSS) or mean-squared-error skill score (MSESS) may seem, it is not well understood and, hence, sometimes misinterpreted. This chapter aims at improving the understanding of the MSESS. We review existing MSESS formulations and classifications, with and without accounting for measurement error. Using simple examples, we illuminate which aspects of prediction quality the MSESS actually measures. It is shown that the MSESS tends to favour model results that underestimate the variance of cumulative bed changes. We further demonstrate that the normalization by the observed cumulative change, which follows from the choice of the initial bed as the reference, is not effective in creating a level playing field over a wide range of prediction situations (trend, episodic event, different seasons). Also, it is shown that the combined presence of larger, persistent scales and smaller, intermittent scales in the cumulative bed changes may lead to an apparent increase of skill with time, while the prediction of neither of these scales becomes more skilful with time. Finally, in order to obtain a balanced appreciation of model performance, the use and development of a more extensive suite of validation measures is advocated.

3.1 Introduction

The introduction of the Brier skill score (BSS) for coastal morphology (Sutherland et al., 2004) was an important step in the further maturing of morphodynamic modelling practice (Roelvink and Reniers, 2012). This BSS essentially is a mean-squared-error skill score (MSESS) measuring the accuracy of a prediction relative to a reference, often the initial bed. Prior to its introduction, the evaluation of the quality of 2D morphological predictions was largely by visual comparison of patterns of sedimentation and erosion between observations and simulations. This is a powerful method, but prone to individual and subjective biases of interpretation. Besides, it is increasingly difficult to apply if there are multiple predictions, as in a sensitivity analysis or ensemble prediction. By yielding a normalized single-number score, the MSESS objectifies the assessment of model performance and allows the intercomparison of quality across a range of prediction situations. Through a generic classification based on its values, predictions receive a quality label. Unsurprisingly, the MSESS has become widely accepted amongst morphodynamic modellers as the preferred way of demonstrating model skill (see also Bosboom et al., 2014, i.e. Ch. 2, and references therein).

A comparative analysis based on skill scores requires a good understanding of the statistics of predictive skill (Gallagher et al., 1998). Along with the MSESS, Sutherland et al. (2004) introduced the Murphey–Epstein decomposition (Murphy and Epstein, 1989) into phase, amplitude, and map-mean error. Although this decomposition can provide valuable insight into specific aspects of prediction quality, it has only been used in a limited number of morphological applications (e.g. van der Wegen et al., 2011; van der Wegen and Roelvink, 2012; Ruessink and Kuriyama, 2008) and seems to be not well understood (Bosboom et al., 2014, i.e. Ch. 2). Further, there have been some accounts of skill scores not matching the researcher's perception of model performance due to the use of the initial bed as the reference (Bosboom et al., 2014; Gallagher et al., 1998; van der Wegen and Roelvink, 2012) or the mean-squared error (MSE) as the accuracy measure (Guerin et al., 2016). Nonetheless, surprisingly little attention has been paid to the interpretation of the MSESS and its values. A recent analysis (Bosboom et al., 2014) of the perception of morphodynamic model skill through the MSESS showed that the apparent simplicity of the MSESS in the context of morphodynamic modelling is deceptive and morphodynamic skill may be misinterpreted as a result.

The main purpose of this chapter is to illustrate the essence of the MSESS using simple, hypothetical examples and examples from literature. We discuss to what extent the MSESS is truly capable of comparing model results for different conditions, regions, time periods et cetera. Further, the question is addressed as to which aspects of prediction quality are exactly measured by the MSESS. In doing so, we also pay attention to formulations and classifications for the MSESS that are adjusted to take measurement errors into account (Sutherland et al., 2004; van Rijn et al., 2003). Section 3.2 summarizes the concept of skill, both in general and in the context of morphodynamic model validation. Next, in Sect. 3.3 due attention is paid to the Murphy–Epstein decomposition (Murphy and Epstein, 1989). Section 3.4 reviews the adjustments to the MSESS (Sutherland et al., 2004; van Rijn et al., 2003) that aim to account for measurement error. The classifications of prediction quality, for the skill scores with and without corrections for measurement error (Sutherland et al., 2004; van Rijn et al., 2003), are discussed in Sect. 3.5. Section 3.6 specifically focusses on the aspects of the MSESS that tend to lead to a misperception of skill. For the reader that is already familiar with the details of the MSESS, this section provides a quick assessment of some common mistakes made in its interpretation. Finally, Sect. 3.7 presents conclusions and discusses strategies for improving model validation efforts.

3.2 What is the Brier skill score?

In this section, we describe the rationale behind the BSS. First, the concept of skill is elaborated on in Sect. 3.2.1. Next, Sect. 3.2.2 presents the definition of the skill score based on the MSE. It has become known amongst coastal modellers as the Brier skill score (BSS), but would more accurately be named MSE skill score

(MSESS) as explained in Sect. 3.2.3. Finally, Sect. 3.2.4 discusses the choice of reference, which is a crucial element of skill.

3.2.1 The concept of skill

The concept of skill refers to the relative accuracy of a prediction compared to a baseline or reference prediction. For a prediction with accuracy E, a skill score ESS can be formulated as follows:

$$\mathrm{ESS} = \frac{E - E_r}{E_\mathrm{i} - E_r} \tag{3.1}$$

where E_r is the accuracy of a baseline or reference prediction and E_i of an impeccable (perfect) prediction. Hence, skill is the difference in accuracy between a prediction and an unskilled reference prediction normalized by the total possible improvement that can be achieved (with respect to the reference prediction). A prediction that is as good as the reference prediction obtains a score of 0 and a perfect prediction a score of 1. A value between 0 and 1 can be interpreted as the proportion of improvement over the unskilled reference prediction. Negative values indicate a prediction worse than the reference prediction.

The concept of skill or relative accuracy aims at making model performance comparable across a range of prediction situations. This allows the ranking of predictions with different (numerical) models and for different geographical locations, forcing conditions, time periods or internal dynamics. A more difficult prediction situation implies a lower accuracy of the reference prediction, such that a lower accuracy is required to obtain a certain skill level. However, in various fields, notably weather forecasting, skill scores were found to not be fully effective in neutralizing the situation's inherent difficulty (Winkler, 1994; Winkler et al., 1996). In Sect. 3.2.4, we will discuss the usual choices of reference, notably the zero change model, for morphodynamic modelling.

3.2.2 Mean-squared-error skill score

For nonprobabilistic predictions of continuous variables, such as wave height or seabed elevation, a common choice for the accuracy measure E in Eq. 3.1 is the MSE. The resulting skill score is often referred to as mean-squared-error skill score outside our field, e.g. Murphy (1988), but is named Brier skill score by coastal modellers following Sutherland et al. (2004)—see Sect. 3.2.3. It reads:

$$\mathrm{MSESS} = \frac{\mathrm{MSE} - \mathrm{MSE}_r}{0 - \mathrm{MSE}_r} = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE}_r} \tag{3.2}$$

since the MSE of a perfect prediction $\mathrm{MSE}_\mathrm{i} = 0$.

The MSE between predictions p and observations o is defined as:

$$\mathrm{MSE} = \left\langle (p - o)^2 \right\rangle = \frac{1}{n} \sum_{i=1}^{n} w_i (p_i - o_i)^2 \tag{3.3}$$

where (p_i, o_i) are the ith pair of the predictions and observations, the angle brackets denote averaging over the n pairs of predictions and observations and w_i are weighting factors with $\sum w_i = 1$. In morphodynamic modelling, the predictand is the bed level, such that p and o are the fields of predicted and observed bed levels z_p and z_o, respectively. Equation 3.2 then operates on the gridded predicted and observed fields by spatially averaging the individual squared differences between the two at each of the n grid points. While for regularly spaced grids $w_i = 1$, for irregularly spaced grids the spatial averaging must be weighted by the grid-cell size represented by w_i. The MSE is known to be unduly sensitive to outliers (Jolliffe and Stephenson, 2012). A clear advantage, however, of using the MSE is that the resulting skill score can readily be decomposed into components that describe specific aspects of prediction quality (see Sect. 3.3). With the MSE according to Eq. 3.3 and the analogous expression for MSE_r, Eq. 3.2 yields:

$$\mathrm{MSESS} = 1 - \frac{\left\langle (p - o)^2 \right\rangle}{\left\langle (r - o)^2 \right\rangle}. \tag{3.4}$$

The MSESS ranges from $-\infty$ to 1 with negative (positive) values indicating that the prediction p is worse (better) than the reference prediction r.

Skill metrics are often formulated in terms of the anomalies (differences) with respect to the reference prediction r. Hence, the anomalies of the predictions and observations are defined as $p' = p - r$ and $o' = o - r$, respectively. In morphodynamic modelling, the anomalies are the predicted and observed sedimentation/erosion differences relative to the reference prediction. In terms of the anomalies, we have for Eq. 3.3:

$$\mathrm{MSE} = \left\langle (p' - o')^2 \right\rangle \tag{3.5}$$

and for the reference prediction r:

$$\mathrm{MSE}_r = \left\langle o'^2 \right\rangle. \tag{3.6}$$

Substitution of Eqs. 3.5 and 3.6 in Eq. 3.2 gives for MSESS in terms of the anomalies (cf. Eq. 3.4):

$$\mathrm{MSESS} = 1 - \frac{\left\langle (p' - o')^2 \right\rangle}{\left\langle o'^2 \right\rangle}. \tag{3.7}$$

The bed levels z_p and z_o are cumulative by nature; they are the sum of the initial bathymetry z_{ini} and the cumulative predicted and observed bed changes from the start of the simulation, Δz_p and Δz_o respectively. Consequently, both the MSE and the MSE$_r$ reflect the cumulative errors from the start of the simulation to the moment of skill evaluation. This is very different from predictands such as wave heights or current magnitudes, which are instantaneous values rather than accumulated quantities over the entire simulation duration. Upon substitution, in Eq. 3.4, of $p = z_{\text{ini}} + \Delta z_p$, $o = z_{\text{ini}} + \Delta z_o$ and $r = z_{\text{ini}} + \Delta z_r$, with Δz_r the cumulative change of the reference prediction relative to the initial bed, we find:

$$\text{MSESS} = 1 - \frac{\left\langle \left(\Delta z_p - \Delta z_o\right)^2 \right\rangle}{\left\langle \left(\Delta z_r - \Delta z_o\right)^2 \right\rangle}. \tag{3.8}$$

Evidently, Eq. 3.8 expresses the MSESS in terms of cumulative (net) changes relative to the initial bed, with the reference prediction yet undefined. Note that from Eq. 3.8 it follows that the MSESS is not altered by the presence of a morphologically inactive region.

Section 3.2.4 deals with the choice of reference in general and the often used reference of zero change in particular. In Sect. 3.6.2, we discuss the implications of the cumulative nature of morphology for the interpretation of morphodynamic model skill through the MSESS.

3.2.3 Naming conventions

The term Brier skill score was first introduced in the field of weather forecasting as the skill of probabilistic dichotomous predictions, using the Brier score (BS) as the accuracy measure (see e.g. Wilks, 2011). Dichotomous events, that is, events that have two mutually exclusive outcomes (e.g. rain or no rain) are generally predicted in terms of probabilities of occurrence. The accuracy of such a probabilistic prediction can be summarized in the Brier score[1]:

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i\right)^2 \tag{3.9}$$

with y_i the forecast probabilities and x_i the observations, which are equal to 1 if the event occurs and equal to 0 if the event does not occur. The BS is essentially the mean-squared error of the probability predictions; it averages the squared differences between pairs of forecast probabilities and binary observations. Hence, the BS can be seen as a special case of the MSE. Like the MSE, a perfect prediction yields BS = 0 and less accurate predictions receive higher scores. Unlike the MSE, the score can only take on values in the range $0 \leqslant \text{BS} \leqslant 1$[2].

[1] What is now universally used as the Brier score (Eq. 3.9) is sometimes more correctly referred to as the half Brier score (see e.g. Wilks, 2011) since it is only half of the score originally introduced by Brier (Brier, 1950).

[2] Corrected from Bosboom and Reniers (2018) where $0 < \text{BS} < 1$ was written.

Now we substitute BS for E in Eq. 3.1, such that the Brier skill score (BSS) reads:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_r} \tag{3.10}$$

where BS is given by Eq. 3.9 and the subscript "r" again refers to the reference prediction. "Our" BSS however—let us refer to it as $\mathrm{BSS}_{\mathrm{cc}}$ with cc representing the coastal community—is an MSE skill score representing the relative accuracy of nonprobabilistic predictions. It is given by the MSESS of Eq. 3.2 rather than the BSS of Eq. 3.10. Within our field, the implications of this misunderstanding upon the introduction of the concept of skill in the coastal community (Sutherland et al., 2004) are limited. However, validation is a generic topic not exclusive to our field and cross-pollination of knowledge and methods between fields will be easier if the naming of generic accuracy and skill measures is consistent. In this paper, we will therefore use the name MSESS for the $\mathrm{BSS}_{\mathrm{cc}}$, well aware though of how established the use of the name BSS in the coastal community is.

3.2.4 Choice of reference

A crucial aspect of skill is the proper selection of the reference. By establishing the zero point at the scale of skill, the reference prediction defines a minimal level of performance and is generally required to be an unskilful, yet not unrealistic forecast as can be made with a naive forecasting method (Winkler, 1994). In weather forecasting, the most appropriate baseline for short-term forecasts is generally assumed to be a persistence forecast, which implies that observations at a certain point are forecast to persist, that is, remain unchanged (Murphy, 1992). For longer-term forecasts, climatology is often considered better, in which case the average of historical data is used as a baseline, such that trends are taken into account (Murphy, 1992).

In morphodynamic modelling, it is common practice to use the initial observed bathymetry at the start of a simulation as the reference ($r = z_{\mathrm{ini}}$), which implies that the model to beat is a model that predicts zero morphological change. The reference of zero morphological change is similar to the concept of persistence in that the initial bed is assumed to persist. However, whereas often a previous state with a constant lag (e.g. the previously observed value) is taken as the persistence reference, the model of zero change is applied irrespective of the prediction horizon by assuming the initial bed to persist. By implication, longer-term trends are not accounted for in the zero change model.

When the reference prediction is taken as the initial bed, such that $\Delta z_r = 0$ in Eq. 3.8, the model to beat is a model of zero morphological change. Now the anomalies in Eq. 3.7 are the cumulative sedimentation/erosion fields from the simulation start time $t = 0$, i.e.: $p' = \Delta z_p$ and $o' = \Delta z_o$. Hence, we find for the MSE skill score through Eq. 3.7:

$$\mathrm{MSESS_{ini}} = 1 - \frac{\left\langle \left(\Delta z_p - \Delta z_o\right)^2 \right\rangle}{\left\langle \Delta z_o^2 \right\rangle} \tag{3.11}$$

with the subscript “ini” referring to the initial bed as the reference prediction. Of course, Eq. 3.11 is also obtained from Eq. 3.8 using $\Delta z_r = 0$. The $\mathrm{MSESS_{ini}}$ can be considered as the fraction of improvement of the model results compared to a model that predicts that no morphodynamic change will occur. It is often interpreted as the model added accuracy relative to a situation in which no modelling is done, other than assuming that the morphology remains unchanged.

Equation 3.11 demonstrates that the use of the zero change model leads to normalization of the error in the bed levels by the observed cumulative change. In other words, at various times in a simulation, the skill is positive as long as the error is bounded by the observed change from the start of the simulation. Hence, in case of larger cumulative changes also larger errors are allowed. The underlying assumption is that the observed cumulative change correctly reflects the intrinsic difficulty of prediction situations (Winkler, 1994; Murphy, 1988; Wilks, 2011; Brier and Allen, 1951) with a different morphological development prior to the evaluation time (for instance trend, cyclic, episodic). In Sect. 3.6.2 we demonstrate, using simple examples, that this cannot be expected to be generally valid, leading to unexpected, counterintuitive results for the $\mathrm{MSESS_{ini}}$.

The zero change model is not the only reference model that can be used. In the presence of a trend, a more appropriate naive model could be some estimate of the trend. For coastline modelling, Davidson et al. (2010) make use of a linear trend prediction as the benchmark, which can be expected to provide a more stringent test than the reference prediction that nothing will change. Especially further in the simulation, the zero change model could become a quite unrealistic benchmark and skill may be overestimated.

Another useful choice of reference is a benchmark prediction with the same model or a different model (Lesser, 2009; Gerritsen et al., 2011)[3]. For such a reference, the issues addressed in Sect. 3.6 can be relevant as well. However, this chapter focusses on the interpretation of the zero change model for morphodynamic model skill.

3.3 Murphy–Epstein decomposition

In this section, we elaborate on the Murphy–Epstein decomposition (Murphy and Epstein, 1989) of the MSESS. First, the equations for the decomposition are given in Sect. 3.3.1. Next, Sect. 3.3.2 provides an overview of 10 hypothetical test cases.

[3] The correct reference is Lesser (2009) rather than Lesser et al. (2004) as mentioned in Bosboom and Reniers (2018).

These are subsequently used, in Sects. 3.3.3 to 3.3.5, to discuss the specific aspects of prediction quality as described by each of the error components.

3.3.1 Decomposition of the MSESS

First, we decompose the MSE, written in terms of the anomalies (Eq. 3.5), as follows (see Bosboom et al., 2014, i.e. Ch. 2, for more details):

$$\mathrm{MSE} = \underbrace{\sigma_{p'}^2 + \sigma_{o'}^2 - 2\sigma_{p'}\sigma_{o'}\rho_{p'o'}}_{\mathrm{MSE_{fluct}}} + \underbrace{\left(\overline{p'} - \overline{o'}\right)^2}_{\mathrm{MSE_{bias}}}. \tag{3.12}$$

Here $\sigma_{p'}$ and $\sigma_{o'}$ are the weighted standard deviations and $\overline{p'}$ and $\overline{o'}$ the weighted map means of p' and o'. Further, $\rho_{p'o'} = \sigma_{p'o'}/(\sigma_{p'}\sigma_{o'})$ is the weighted Pearson correlation coefficient between p' and o', with $\sigma_{p'o'}$ representing the weighted covariance. Note that the MSE consists of a part that expresses the mismatch between the fluctuating parts in predictions and observations ($\mathrm{MSE_{fluct}}$) and a bias part that quantifies the systematic error or map-mean error ($\mathrm{MSE_{bias}}$).

For the reference prediction, we deduce from Eq. 3.12 (cf. Eq. 3.6):

$$\mathrm{MSE}_r = \sigma_{o'}^2 + \overline{o'}^2 \tag{3.13}$$

with on the right side first $\mathrm{MSE}_{r,\mathrm{fluct}}$, and then $\mathrm{MSE}_{r,\mathrm{bias}}$. Rearrangement of the terms in Eqs. 3.12 and 3.13 gives:

$$\mathrm{MSE} = \sigma_{o'}^2\left(1 - \alpha' + \beta' + \gamma'\right) \ \text{ with} \tag{3.14}$$

$$\alpha' = \rho_{p'o'}^2 \tag{3.15}$$

$$\beta' = \left(\rho_{p'o'} - \frac{\sigma_{p'}}{\sigma_{o'}}\right)^2 \tag{3.16}$$

$$\gamma' = \frac{\left(\overline{p'} - \overline{o'}\right)^2}{\sigma_{o'}^2} \tag{3.17}$$

and

$$\mathrm{MSE}_r = \sigma_{o'}^2\left(1 + \epsilon'\right) \ \text{ with} \tag{3.18}$$

$$\epsilon' = \frac{\overline{o'}^2}{\sigma_{o'}^2}. \tag{3.19}$$

Finally, the substitution of Eqs. 3.14 and 3.18 in Eq. 3.2 yields the Murphy–Epstein decomposition (Murphy and Epstein, 1989) of the MSE skill score:

$$\text{MSESS} = \frac{\alpha' - \beta' - \gamma' + \epsilon'}{1 + \epsilon'}. \tag{3.20}$$

The terms α', β', γ' are often referred to as phase association, conditional bias or amplitude error and systematic bias or map-mean error, respectively, whereas ϵ' has been explained as an adjustment due to the observed map-mean (Livezey et al., 1995). In the next section, we investigate what exactly is measured by these terms. In doing so, we will refer to the linear least-squares regression of o' given p':

$$o' = b_0 + b_1 p' \text{ with } b_1 = \rho_{p'o'} \frac{\sigma_{o'}}{\sigma_{p'}} \text{ and } b_0 = \overline{o'} - b_1 \overline{p'} \tag{3.21}$$

where b_1 and b_0 are the slope and the o'-intercept of the regression line, respectively. The correlation coefficient $\rho_{p'o'}$ $(-1 \leqslant \rho_{p'o'} \leqslant 1)$ reflects the overall strength and direction of the linear correspondence; scatter around the best linear fit leads to a magnitude smaller than 1, with the sign indicating positive or negative correspondence and, hence, slope b_1. With Eq. 3.21, we can relate β', given by Eq. 3.16, to the slope b_1 as follows:

$$\beta' = \frac{\sigma_{p'}^2}{\sigma_{o'}^2} (b_1 - 1)^2. \tag{3.22}$$

3.3.2 Overview of test cases

In Sects. 3.3.3 to 3.3.5, the meaning of the error components α', β', γ' and ϵ' is investigated using 10 cases (P1–P10; Table 3.1). For each of these cases, the observations o' (solid lines in Figs. 3.1 to 3.3) consist of a distinct sedimentation and erosion feature indicated by positive and negative values for o', respectively. The domain-averaged bed change $\overline{o'} = 0$. The error values of the predictions P1–P10, which vary with respect to $\rho_{p'o'}$, $\sigma_{p'}/\sigma_{o'}$ and/or map-mean error $\overline{p'} - \overline{o'}$, are listed in Table 3.1 and will be explained in the following sections.

3.3.3 Structural similarity α'

The term α' (Eq. 3.15) indicates the tendency of o' and p' to vary together and hence expresses the similarity in the structure of the sedimentation/erosion fields $(0 \leqslant \alpha' \leqslant 1)$. In terms of a regression model, α' is the proportion of the variation in the values of o' that can be linearly explained by p' (and vice versa). Similarly, the term $1 - \alpha'$ is the unexplained proportion of the variance (cf. Eq. 3.14). A deviation of α' from 1 indicates scatter around the best linear fit, which could be indicative of incorrect positions (Fig. 3.1, left panel), relative magnitudes (Fig. 3.1,

	Error values							
Case	$\rho_{p'o'}$	$\sigma_{p'}/\sigma_{o'}$	b_1	α'	β'	γ'	ϵ'	MSESS
P1	0.60	1.00	0.60	0.37	0.16	0.00	0.00	0.21
P2	0.82	0.67	1.23	0.68	0.02	0.06	0.00	0.60
P3	0.93	0.74	1.25	0.86	0.03	0.00	0.00	0.82
P4	1.00	1.00	1.00	1.00	0.00	0.44	0.00	0.56
P5	1.00	1.00	1.00	1.00	0.00	0.44	0.00	0.56
P6	1.00	0.60	1.67	1.00	0.16	0.00	0.00	0.84
P7	−1.00	1.00	−1.00	1.00	4.00	0.00	0.00	−3.00
P8	0.50	1.00	0.50	0.25	0.25	0.00	0.00	0.00
P9	0.50	1.00	0.50	0.25	0.25	0.00	0.00	0.00
P10	0.60	0.60	1.00	0.37	0.00	0.00	0.00	0.37

Table 3.1: Ten predicted anomalies p' compared to the same observed anomaly o' (solid lines in Figs. 3.1 to 3.3).

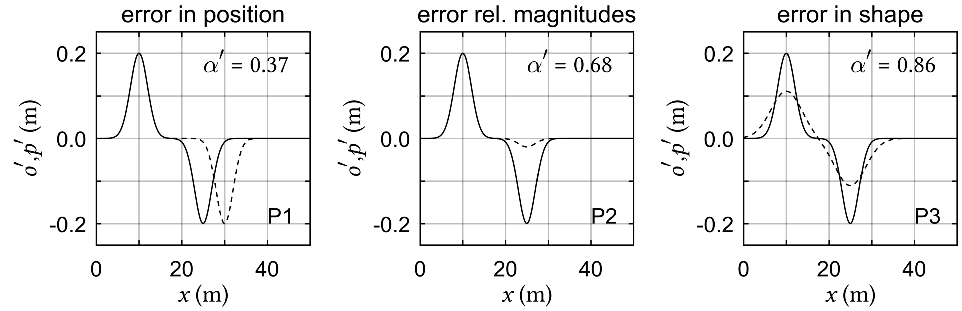


Figure 3.1: The structural similarity α' deviates from 1 in the case of errors in position (*left*), relative magnitudes (*middle*) or shape (*right*) of the sedimentation/erosion features. Solid lines: observations; dashed lines: predictions (prediction numbers in accord with Table 3.1).

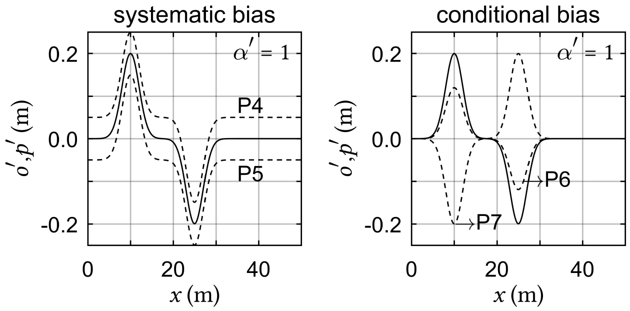


Figure 3.2: Systematic bias (*left*) and conditional bias (*right*) do not influence α'. Solid lines: observations; dashed lines: predictions (prediction numbers in accord with Table 3.1).

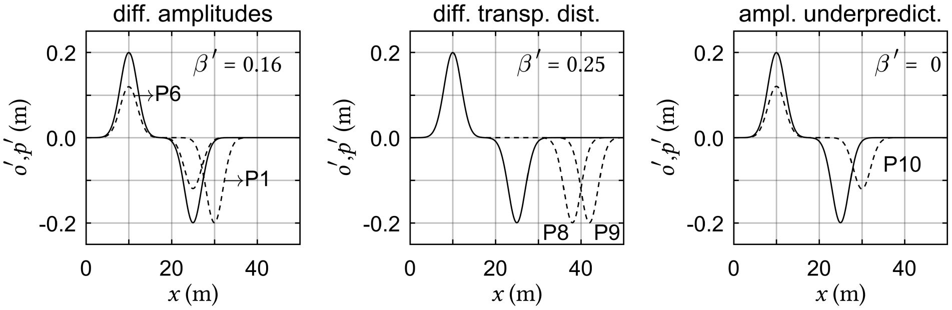


Figure 3.3: Conditional bias β' for pairs of observations (solid lines) and predictions (dashed lines, numbers in accord with Table 3.1). *Left*: P1 and P6 have the same β', but the predicted standard deviation $\sigma_{p'}$ is different; *middle*: P8 and P9 have the same β', but the transport distance between o' and p' is different; *right*: P10 is perfect in terms of β', but $\sigma_{p'}$ is too small.

middle panel) or shapes (Fig. 3.1, right panel) of features in the (de-meaned) sedimentation/erosion fields. Structural dissimilarity is therefore a more inclusive description of $1 - \alpha'$ than the names phase error, suggested in Livezey et al. (1995), or position error, as employed in Gerritsen et al. (2011). Likewise, the explanation of Sutherland et al. (2004) that a deviation of α' from 1 means that sand is moved to the wrong position may be too restrictive; erosion and sedimentation features that are at the right location but have, for instance, incorrect relative magnitudes also give $\alpha' < 1$ (Fig. 3.1, middle panel).

The value of α' is not sensitive to biases that may be present in the predictions, that is, when the predictions are changed with a (positive or negative) constant (Fig. 3.2, left panel) or constant factor (Fig. 3.2, right panel), the value of α' remains the same. As a consequence, $\alpha' = 1$ could mean that the fields are opposite in sign (P7; see Table 3.1 and Fig. 3.2, right panel). Hence, the *direction* of the linear correspondence can only be determined by evaluating $\rho_{p'o'}$ instead of α'.

3.3.4 Scale error β'

The error term β' (Eq. 3.16) is the penalty due to conditional bias ($0 \leqslant \beta' < \infty$) with $\beta' = 0$ indicating no bias. Following Livezey et al. (1995), Sutherland et al. (2004) have explained the conditional bias β' as an amplitude error indicating that "the wrong volumes of sand have been moved." This interpretation must be used with care, as demonstrated by Fig. 3.3 (left panel). Here, case P6, which has a perfect correlation ($\rho_{p'o'} = 1$), receives a nonperfect $\beta' = 0.16$, due to an underestimation of the variation in bed changes ($\sigma_{p'}/\sigma_{o'} = 0.6$; see Table 3.1). For case P1, however, the same value of $\beta' = 0.16$ is obtained, even though the amplitudes of the bed changes are perfectly modelled, viz. $\sigma_{p'}/\sigma_{o'} = 1$. Now, the nonperfect value of β' is due to a deviation of the correlation from 1 ($\rho_{p'o'} = 0.6$; see Table 3.1). Clearly,

β' is not a proper indicator of the extent to which the amplitudes of bed changes are correctly modelled.

The interpretation of Gerritsen et al. (2011), who explain β' as a transport error, should be treated with similar care. This can be seen from P8 and P9 (Fig. 3.3, middle panel), which both have $\beta' = 0.25$ (see Table 3.1). Nonetheless, the domain-integrated transport errors are different, as can be deduced from the distance between the predicted and observed location of erosion being larger for P9 than for P8. Other examples are provided by P4 and P5 (Fig. 3.2, left panel). These predictions obtain $\beta' = 0$, even though the errors in bed levels throughout the domain indicate incorrect sediment transport rate gradients.

How, then, should β' be interpreted? From Eq. 3.22, it follows that, for nonzero $\sigma_{p'}$, $\beta' = 0$ implies $b_1 = \rho_{p'o'}\sigma_{o'}/\sigma_{p'} = 1$, i.e. a 1:1 slope of the regression line (Eq. 3.21). A nonzero value for β' indicates that b_1 deviates from 1. In the absence of map-mean errors, a slope $b_1 > 1$ is equivalent to an underprediction, by the regression model, of the larger values and an overprediction of the smaller values. For $b_1 < 1$, the regression model underestimates the smaller values and overestimates the larger values. Deviations of b_1 from 1 can occur due to a deviation of $\sigma_{o'}/\sigma_{p'}$ and/or $\rho_{p'o'}$ from 1 (see Table 3.1).

In terms of a regression analysis, it is desirable to have $b_1 = 1$, which for positive, nonperfect correlation ($0 < \rho_{p'o'} < 1$) can be achieved by scaling the magnitude of the predicted anomalies to $\sigma_{p'} = \rho_{p'o'}\sigma'_o$ (P10 in Fig. 3.3, right panel). For negative correlation, the conditional bias is minimized for $\sigma_{p'} = 0$. Optimizing β' thus implies an underestimation of the variance of the anomalies. Compare for instance cases P1 (Fig. 3.1, left panel) and P10 (Fig. 3.3, right panel), which have the same suboptimal anomaly correlation ($\rho_{p'o'} = 0.6$; Table 3.1). A lower β' and a higher MSESS are found for P10 (with $\sigma_{p'}/\sigma_{o'} = 0.6$) than for P1 (with $\sigma_{p'}/\sigma_{o'} = 1$).

In conclusion, a nonzero value of β' indicates a suboptimal scaling of the magnitude of the anomalies to account for the value of the correlation. “Optimal” is defined here in terms of the smallest overall least-squares error and skill score and not in terms of the variance of the bed changes, which must be judged separately (see Sect. 3.6.1). Sutherland et al. (2004) implicitly acknowledge the ambiguity about the term amplitude error for β' by stating that perfect modelling of phase (represented by $\rho_{p'o'}$) and amplitude (represented by $\sigma_{p'}/\sigma_{o'}$) gives $\beta' = 0$. The opposite however is not true: $\beta' = 0$ does not imply that $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$ are modelled perfectly. In Sect. 3.6.1, we will further investigate the consequences for morphodynamic modelling.

3.3.5 Bias terms γ' and ϵ'

The error term γ' (Eq. 3.17) is a normalized map-mean error or sediment budget error ($0 \leqslant \gamma' < \infty$). A nonzero value indicates a misestimation of the amount of sediment imported into or exported from the model domain and hence of the average bed level (Gerritsen et al., 2011). Such a systematic bias, as for P4 and P5 (Fig. 3.2, left panel) offsets the intercept of the regression line of o' given p' without changing the slope of the line (Eq. 3.21). In order to establish the direction of the misestimation, (the sign of) $\overline{p'} - \overline{o'}$ must be considered.

Similarly, ϵ' (Eq. 3.19) is the normalized map-mean error of the reference prediction referred to by Livezey et al. (1995) as a normalization error ($0 \leqslant \epsilon' < \infty$). A nonzero value indicates a map-mean sediment budget (and hence bed level) difference between the reference prediction and the observations. For larger ϵ', the reference becomes a worse prediction with respect to the mean.

For the interpretation of γ' and ϵ' for morphological models, it is important to realize that both error terms are normalized by the variance of the observed cumulative change away from the reference prediction. Compare for instance case P5 (Fig. 3.2, left panel) and case R1, the latter differing from P5 only in that o' and p' are reversed. In both cases, the same absolute inaccuracy develops (α', β', γ' being identical; Table 3.2). However, for R1, the mean of the observations shows a trend that is not followed by the predictions, whereas for P5, the predictions have a trend that is not observed. The latter situation gives a lower skill due to the smaller normalization error ϵ'. The higher skill in the first case is due to the reference prediction being a worse predictor when the observations develop away from the reference. The implications are further analyzed in Sect. 3.6.2.

		Error values							
Case	Description	$\rho_{p'o'}$	$\sigma_{p'}/\sigma_{o'}$	b_1	α'	β'	γ'	ϵ'	MSESS
P5	Systematic bias	1.00	1.00	1.00	1.00	0.00	0.44	0.00	0.56
R1	o' and p' reversed	1.00	1.00	1.00	1.00	0.00	0.44	0.44	0.69

Table 3.2: Effect of systematic bias for P5 and the same case with o' and p' reversed.

3.4 Adjusted formulations with measurement error

Equation 3.2 assumes that the MSE of a perfect prediction is zero, which means that the presence of errors in the data is not taken into account. In reality, data errors will occur due to errors in the bathymetric surveys as well as in the subsequent interpolation procedure to a common grid. As a consequence, the deviation from 1 of the skill value according to Eq. 3.2 is not purely due to prediction error, but can partly be attributed to data errors. Two methods have been proposed to take the effect of the latter out of the equation.

The first method (Sutherland et al., 2004) assumes that the initial and final measured bathymetries consist of an actual bathymetry and independent, random measurement errors δ with the same $\delta_{\mathrm{rms}} = \sqrt{\langle \delta^2 \rangle}$. The $\mathrm{MSE_i}$ of the perfect prediction is assumed to be equal to $\mathrm{MSE_i} = 2\langle \delta^2 \rangle$ made up of a part $\langle \delta^2 \rangle$ due to the error in the final bathymetry to which the perfect prediction is compared and another part $\langle\delta^2\rangle$ that is present in the initial bathymetry. This initial error is assumed to persist throughout the simulation. Now Eq. 3.1 leads to:

$$\mathrm{MSESS}_{\mathrm{ini,\,P}} = \frac{\mathrm{MSE}_r - \mathrm{MSE}}{\mathrm{MSE}_r - 2\langle\delta^2\rangle} \tag{3.23}$$

such that a prediction already obtains perfect skill with $\mathrm{MSE} = 2\langle\delta^2\rangle$ instead of $\mathrm{MSE} = 0$. Also, for the same MSE and MSE_r, positive skill scores increase but negative ones decrease. The denominator of Eq. 3.23 can be seen as the actual MSE of the reference prediction, $\mathrm{MSE}_{r,\mathrm{a}}$, after correction for measurement errors in both the final bathymetry and the initial bathymetry: $\mathrm{MSE}_r - 2\langle\delta^2\rangle = \mathrm{MSE}_{r,\mathrm{a}}$.

To demonstrate the effect of measurement error, we elaborate on an example for East Pole Sand, reported in Sutherland et al. (2004). Three predictions, which only differ with respect to the values of the median grain diameter d_{50}, are compared relative to the same initial bed as the reference (see Table 3.3[4]). As a consequence, the MSE_r is the same across the simulations, but the MSE varies, being smaller for larger d_{50}. For each of the simulations, the $\mathrm{MSESS}_{\mathrm{ini}}$ is computed for $\delta_{\mathrm{rms}} = \sqrt{\langle\delta^2\rangle} = 0$, 0.026 and 0.1 m, respectively. The table shows that for larger δ_{rms}, the denominator $\mathrm{MSE}_r - 2\langle\delta^2\rangle$ decreases and the (positive) skill increases, since the distance from the reference to a perfect prediction is $2\langle\delta^2\rangle$ shorter.

d_{50} (mm)	MSE (m^2)	MSE_r (m^2)	δ_{rms} (m)	$\mathrm{MSE}_r - 2\langle\delta^2\rangle$ (m^2)	$\mathrm{MSESS}_{\mathrm{ini,P}}$	$\mathrm{MSESS}_{\mathrm{ini,vR}}$
0.25	0.0317	0.037	0.000	0.037[5]	0.15	0.15
0.25	0.0317	0.037	0.026	0.036	0.15	0.30
0.25	0.0317	0.037	0.100	0.017	0.31	0.59
0.35	0.0266	0.037	0.000	0.037	0.29	0.29
0.35	0.0266	0.037	0.026	0.036	0.30	0.42
0.35	0.0266	0.037	0.100	0.017	0.61	0.68
0.50	0.0246	0.037	0.000	0.037	0.34	0.34
0.50	0.0246	0.037	0.026	0.036	0.35	0.47
0.50	0.0246	0.037	0.100	0.017	0.74	0.71

Table 3.3: Skill scores for East Pole Sand accounting for measurement error according to Eqs. 3.23 and 3.25. These reduce to Eq. 3.2 for $\delta_{\mathrm{rms}} = 0$.

[4] Values in Table 3.3 are partly taken directly (columns 1, 4, 6 and 7) and partly deduced (columns 2, 3 and 5) from Sutherland et al. (2004). In Bosboom and Reniers (2018) on the contrary, column 6 listed $\mathrm{MSESS}_{\mathrm{ini,P}}$ values for $\delta_{\mathrm{rms}} \neq 0$ that were computed from Eq. 3.23, leading to minor roundoff differences compared to the here reported values.

[5] Values in this column are corrected from Bosboom and Reniers (2018) where by mistake values were reported corresponding to $\mathrm{MSE}_r - \langle\delta^2\rangle$ rather than $\mathrm{MSE}_r - 2\langle\delta^2\rangle$. This is without further consequences since the correct values were used to compute $\mathrm{MSESS}_{\mathrm{ini,P}}$.

Using Eq. 3.23 to compute skill values is equivalent to using Eq. 3.2 with actual $\mathrm{MSE_a}$ values, which are $2\langle\delta^2\rangle$ lower than the full MSE values due to the presence of errors in the initial and final bathymetries. This can be verified by substitution of $\mathrm{MSE} = \mathrm{MSE_a} + 2\langle\delta^2\rangle$ and $\mathrm{MSE}_r = \mathrm{MSE}_{r,\mathrm{a}} + 2\langle\delta^2\rangle$ in Eq. 3.23. We find:

$$\mathrm{MSESS_{ini,P}} = \frac{\mathrm{MSE}_{r,\mathrm{a}} - \mathrm{MSE_a}}{\mathrm{MSE}_{r,\mathrm{a}}} = 1 - \frac{\mathrm{MSE_a}}{\mathrm{MSE}_{r,a}} \tag{3.24}$$

or the “normal” skill formulation Eq. 3.2 applied to that part of the MSE errors that can be attributed to the predictions. Hence, as long as the actual errors of the reference prediction and prediction remain unchanged, also the skill values remain unchanged. This is further illustrated by an example described in Sect. 3.5.2 (see Fig. 3.4).

Van Rijn et al. (2003) use a different approach by adjusting Eq. 3.2 only with respect to MSE in the numerator of the second term to the right:

$$\mathrm{MSESS_{ini,vR}} = 1 - \frac{\left\langle \left(\max\left(|p' - o'| - \delta, 0\right)\right)^2 \right\rangle}{\mathrm{MSE}_r}. \tag{3.25}$$

In other words, per grid point only the part of the error is considered that is larger than an error δ for which we assume, based on Sutherland et al. (2004), that δ_{rms} must be taken. If the error is smaller than δ_{rms}, the error is set to zero. Comparison with Eq. 3.24 shows that the error term in the numerator may be seen as van Rijn’s formulation for $\mathrm{MSE_a}$. Thus, according to van Rijn $\mathrm{MSE} - \mathrm{MSE_a} \leqslant \langle\delta^2\rangle$. However, since the reference prediction in the denominator is not corrected, the skill values increase in the presence of measurement error, even if the actual errors of the reference prediction and prediction remain unchanged (see Fig. 3.4). Consequently, the method of van Rijn et al. (2003) generally gives the larger improvement in skill score for the same error (Table 3.3 and Sutherland et al., 2004).

As already advocated by Sutherland et al. (2004), Eq. 3.23 is the recommended formulation to take measurement error into account, since as opposed to Eq. 3.25 it is consistent with the definition of skill. Note, however, that alongside these adjusted formulations for measurement error, adjusted rankings have been proposed. Their validity is discussed in Sect. 3.5.2.

3.5 Generic ranking of model results

A prediction’s skill score indicates the proportion of improvement over a baseline or reference prediction. Let us suppose that the normalization by the change relative to the reference prediction is able to create a level playing field. We can now directly compare skill scores from different prediction situations as well as classify predictions based on their skill score. In this section, we discuss the classifications as have been suggested in morphodynamic modelling for the MSE skill score (Sect. 3.5.1) and the variations hereon to account for measurement error (Sect. 3.5.2). These classifications aim to provide a qualitative judgement of model quality. We will show that the skill score formulations and classifications are not independent and that in literature the rankings are often used inconsistently.

3.5.1 Ranking in absence of measurement error

Sutherland et al. (2004) propose a classification of model quality based on the scores of the $\mathrm{MSESS}_{\mathrm{ini}}$ as given by Eq. 3.2 using the initial bed as the reference prediction (see Table 3.4). As a lower limit of a useful (i.e. good) prediction, a value of $\mathrm{MSESS}_{\mathrm{ini}} = 0.2$ is suggested. Sutherland et al. (2004) refer to a practice in atmospheric sciences of considering an anomaly correlation coefficient of 0.6 as a lower limit of a useful medium range forecast (e.g. Hollingsworth et al., 1980). With $\gamma' = \epsilon' = 0$ and $\sigma_{p'}/\sigma_{o'} = 1$, this leads to $\mathrm{MSESS}_{\mathrm{ini}} = 2\rho_{p'o'} - 1 = 0.2$.

Nonetheless, in atmospheric sciences a skill of 0.2 is not generally considered a (lower limit of a) good skill score. For instance, Murphy and Epstein (1989) note that a 60% level for the correlation coefficient means that only 36% of the variation in observations is explained by the variation in the computations and that when errors in biases and scale (ignored in the correlation coefficient) are taken into account, the MSE skill value will be lower, for example, 20% as above. They rightfully point out that this is only 20% rather than 60% on the way of a perfect forecast, which is a quite low value to be labelled as "good". Of course, there is no fixed rule for determining what skill is considered excellent, good, moderate, poor or bad and one can imagine that this is also dependent on the (state of the modelling in) the field under consideration. In the interpretation of Table 3.4, it is good to keep in mind that the labelling is quite forgiving, probably in line with the present state of morphodynamic modelling. In morphodynamic modelling, values for the anomaly correlation coefficient, the correlation coefficient between computed and measured cumulative sedimentation/erosion patterns, rarely exceed 0.6. This can be different for hydrodynamic model skill (Baart et al., 2016).

	$\mathrm{MSESS}_{\mathrm{ini}}$	$\mathrm{MSESS}_{\mathrm{ini,P}}$	$\mathrm{MSESS}_{\mathrm{ini,vR}}$
Excellent	1.0−0.5	1.0−0.8	1.0−0.8
Good	0.5−0.2	0.8−0.3	0.8−0.6
Reasonable/fair	0.2−0.1	0.3−0.15	0.6−0.3
Poor	0.1−0.0	0.15−0.0	0.3−0.0
Bad	<0.0	<0.0	<0.0

Table 3.4: Classification according to Sutherland et al. (2004) for the MSE skill score as in Eqs. 3.2, 3.23 and 3.25.

Sutherland et al. (2004) noted that as skill scores are used more and more, it would become apparent what value of skill score is needed before a model result can be considered good. However, such a fine-tuning has hardly taken place and, instead, skill scores are sometimes combined with alternative classifications without further discussion. For instance, $\mathrm{MSESS_{ini}}$ according to Eq. 3.2 is sometimes judged (see e.g. Baart et al., 2016; El kadi Abderrezzak and Paquier, 2009) based on the classification proposed by van Rijn et al. (2003), see Table 3.4, which raises the bar for a good morphological prediction from $\mathrm{MSESS_{ini}} = 0.2$ to 0.6. Baart et al. (2016) consider the volume of intertidal change as the persisted variable, instead of fields of bathymetric change, such that a lower limit of 0.6 for a good forecast may certainly be defendable. Nonetheless, the reference to van Rijn's classification (van Rijn et al., 2003) seems less appropriate, it being specifically intended to match the skill score $\mathrm{MSESS_{ini,vR}}$ (Eq. 3.25) for fields of bathymetric change in the presence of measurement error (see Sect. 3.5.2).

3.5.2 Ranking in case of measurement error

The alternative skill formulations presented in Sect. 3.4 aim to correct for the influence of measurement error, such that only prediction error is penalized. In addition, Sutherland et al. (2004) and van Rijn et al. (2003) suggest that these adjusted skill formulations should be evaluated using alternative classifications for the skill scores, as found in the third and fourth columns of Table 3.4 for the skill defined by Eqs. 3.23 and 3.25, respectively. Conceptually, however, a skill formulation that is effective in removing measurement error should yield values that can be directly compared to the classification valid in the absence of data errors (Sect. 3.4). We will illuminate the above using an artificial case of the formation of a rip channel.

First, recall that the measured initial and final bathymetries are assumed to be the sum of an actual bathymetry and fields of random measurement error. Hence, for the zero change reference, we have $r = r_\mathrm{a} + \delta_\mathrm{ini}$ and for the observations $o = o_\mathrm{a} + \delta_\mathrm{final}$ with $\sqrt{\langle \delta_\mathrm{ini}^2 \rangle} = \sqrt{\langle \delta_\mathrm{final}^2 \rangle} = \delta_\mathrm{rms}$ (Sect. 3.4). Also, equivalent to the perfect prediction according to Sutherland et al. (2004), we assume $p = p_\mathrm{a} + \delta_\mathrm{ini}$ (Sect. 3.4). In our example, a planar beach serves as the initial bathymetry r_a. Next, the beach is modified with two artificial rip channels, slightly different in shape and position. The one morphology serves as the observations (o_a) and the other as the predictions (p_a). The corresponding mean-squared errors for $\delta_\mathrm{rms} = 0$ are $\mathrm{MSE_a} = 0.051\,\mathrm{m}^2$ and $\mathrm{MSE}_{r,\mathrm{a}} = 0.073\,\mathrm{m}^2$ and the corresponding $\mathrm{MSESS_{ini}} = 0.31$. Next, we add noise fields to the fields of r, o and p and compute the skill values according to Eq. 3.2 and Eqs. 3.23 and 3.25 as a function of δ_rms (Fig. 3.4). The figure shows both the expected values of the skill scores and the standard deviation bands. Since MSE and MSE_r are now larger than $\mathrm{MSE_a}$ and $\mathrm{MSE}_{r,\mathrm{a}}$, respectively, by $2\langle \delta^2 \rangle$, the skill according to Eq. 3.2 decreases with increasing measurement error, such that the predictions are unjustly penalized for the presence of measurement error. The adjusted skill formulation according to Sutherland et al. (2004), Eq. 3.23, shows a constant skill of 0.31, in line with the fact that the actual predicted and measured bathymetries are unchanged. Now the addition of measurement error does not change the skill score, showing that Eq. 3.23 is effective in removing the influence of the measurement error. This was also demonstrated by the analysis in Sect. 3.4 leading to Eq. 3.24. The constant skill irrespective of the measurement error indicates that an adjusted classification of the skill score is not appropriate; skill scores according to Eq. 3.23 should be judged using the same classification as for Eq. 3.2 valid without measurement error. This is further underlined by the fact that Eq. 3.23—as well as Eq. 3.25—reduces to Eq. 3.2 for $\delta_{\mathrm{rms}} = 0$.

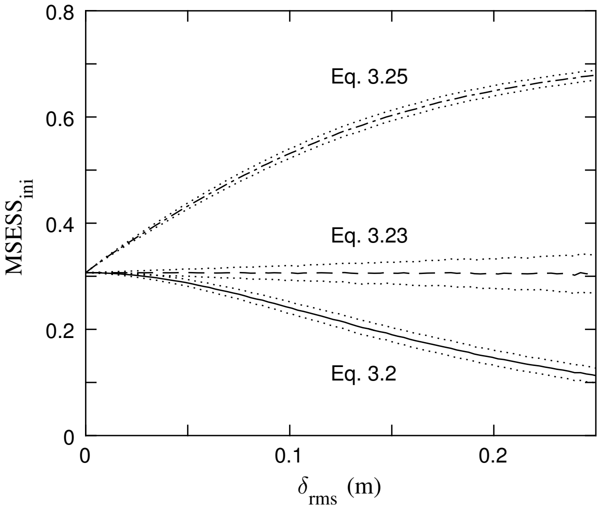


Figure 3.4: Effect of the addition of measurement error on the skill scores according to Eqs. 3.2, 3.23 and 3.25; expected values of the skill scores and standard deviation bands.

Van Rijn's adjusted formulation (van Rijn et al., 2003), Eq. 3.25, gives a strong increase of the skill score (Fig. 3.4), due to the fact that there is no correction for measurement error in the denominator of the second term of the right-hand-side of Eq. 3.25. In addition, the adjusted MSE in the numerator does not consistently remove the effect of measurement error from the MSE, but was seen to vary with δ_{rms}. The suggestion to adjust the skill ranking can be understood from the inflation of the skill scores by adding measurement error. However, the mapping of the skill formulations and their rankings was performed for the specific situation of East Pole Sand (Table 3.3) and for $\delta_{\mathrm{rms}} = 0.1$ m only (Sutherland et al., 2004). A universal mapping cannot be made since the required adjustment depends on the measurement error and the situation.

In conclusion, when a correction for measurement error is called for, we advise the use of the skill formulation according to Sutherland et al. (2004), i.e. Eq. 3.23, rather than van Rijn et al. (2003), i.e. Eq. 3.25, in combination with a classification that is not adjusted for measurement error.

3.6 Three common misinterpretations of skill

The use of the MSESS for morphodynamic predictions requires that the MSE is an appropriate measure of correspondence and that the cumulative change away from the reference correctly reflects the inherent ease or difficulty of the prediction situation. Therefore, we will first discuss, in Sect. 3.6.1, the consequences of the use of the MSE. Subsequently, Sect. 3.6.2 pays attention to the normalization by the cumulative change away from the reference. The effect of the presence of multiple spatial scales is addressed in Sect. 3.6.3.

3.6.1 Smooth is better

The MSE and other overall point-wise metrics are prone to penalize rather than reward the correct prediction of variability[6] (Anthes, 1983; Taylor, 2001). As a consequence, featureless predictions are sometimes favoured over predictions whose features are misplaced, a characteristic that is referred to as “double penalty effect” (Bougeault, 2003). From the perspective of coastal morphologists, this may lead to wrong conclusions concerning the ranking of predictions.

The double penalty effect is inherited by the MSESS resulting in a tendency to reward the underestimation of the variance of morphodynamic change. As a consequence, predictions of sedimentation/erosion features that are correct in terms of magnitude but are misplaced in space may not outperform even the reference prediction of zero change. This is nicely illustrated by a numerical hindcast of morphological changes of a wide estuary mouth sandbank located along the French Atlantic Coast (Guerin et al., 2016). Of two morphodynamic simulations, the simulation that captures several of the main morphological changes receives a lower score ($\mathrm{MSESS_{ini}} = -0.18$) than the simulation that predicted almost no morphological change ($\mathrm{MSESS_{ini}} = 0.01$). Evidently, the latter prediction is close to the reference prediction of zero change and, thus, $\mathrm{MSE} \approx \mathrm{MSE}_r$ and $\mathrm{MSESS_{ini}} \approx 0$. The prediction that reproduced some important features receives a larger MSE and hence a negative skill.

The tendency to underestimate the magnitude of bed changes[7] is easily demonstrated through the behaviour of β' (Sect. 3.3.4). For positive anomaly correlation, β' is minimized and hence the skill is maximized for $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'}$ (P10; Fig. 3.3, right panel). If the variance of the bed changes is predicted correctly, i.e. $\sigma_{p'}/\sigma_{o'} = 1$ (P1; Fig. 3.1, left panel), a lower skill would be reported (see Table 3.1). This runs contrary to the intuitive idea of having optimal performance when $\sigma_{p'} = \sigma_{o'}$.

[6] Here, we implicitly assume that location errors are inevitable and define variability as the standard deviation or variance of the bed levels or bed changes at the scales of interest (addendum to Bosboom and Reniers, 2018).

[7] Or, rather, the tendency of the $\mathrm{MSESS_{ini}}$ *to reward* the underestimation of the magnitude of bed changes (addendum to Bosboom and Reniers, 2018).

A real-life illustration is provided for by the comparison of observed and predicted bathymetric changes for East Pole Sand (Sutherland et al., 2004), which was already referred to in Sect. 3.4. From Table 3.5, it can be seen that increasing the d_{50} does not affect the values for the anomaly correlation $\rho_{p'o'}$ (and hence α'). Also, γ' and ϵ' do not differ significantly for the three predictions. A larger d_{50}, however, strongly reduces $\sigma_{p'}/\sigma_{o'}$. The smallest MSE and largest $\text{MSESS}_{\text{ini}}$ are achieved by the prediction that shows the most severe underprediction of the variance of the bed changes. Clearly, when maximizing the MSESS is the objective, underpredicting the variability of the bed changes would be advantageous.

d_{50} (mm)	MSE (m^2)	$\text{MSESS}_{\text{ini}}$	α'	β'	γ'	ϵ'	$\rho_{p'o'}$	$\sigma_{p'}/\sigma_{o'}$
0.25	0.0317	0.15	0.38	0.20	0.04	0.01	0.62	1.06
0.35	0.0266	0.29	0.38	0.07	0.03	0.01	0.62	0.88
0.50	0.0246	0.34	0.38	0.01	0.03	0.01	0.62	0.52/0.72

Table 3.5: Error values for East Pole sand (partly taken and partly deduced from Sutherland et al., 2004).

As a last example, we refer to the morphodynamic simulations of the Bornrif (Achete et al., 2011), a dynamic attached bar at the Wadden Sea island of Ameland, which we analyzed in Bosboom et al. (2014), i.e. Ch. 2. Throughout the 15 years of simulation $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$ are close together in value with in the last year (2008) $\rho_{p'o'} \approx \sigma_{p'}/\sigma_{o'} \approx 0.66$. Since γ' and ϵ' decrease towards zero during the simulations, the skill in 2008 is equal to the so-called potential skill in the absence of biases $\alpha' = \rho^2_{p'o'} = 0.45$. For $\sigma_{p'}/\sigma_{o'} = 1$, however, the skill would have been lower with $\beta' = (0.66 - 1)^2 = 0.12$ at 0.33.

In modelling practice, a reduction of the overall size of bed changes is easier to achieve, for instance by changing the grain size or a transport parameter, than an improvement of the anomaly correlation coefficient. It may therefore well be that in many modelling studies, without the modeller necessarily being aware of this, the ratio of predicted over observed anomaly standard deviation is lowered towards the level of the correlation. Although this certainly optimizes the $\text{MSESS}_{\text{ini}}$, another aspect of model quality, the variance of bed changes, is less well predicted. We therefore advocate that values of $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$ are explicitly reported and a deliberate choice is made for the "optimal" simulation. Clearly, this finding also has implications for (automated) calibration procedures that minimize MSE or $\text{MSESS}_{\text{ini}}$.

3.6.2 Cumulative versus absolute change

The essential assumption underlying the concept of skill is that the reference prediction provides a fair normalization of the accuracy across a range of prediction situations (choice of model, length of the simulation, geographical locations, forcing conditions, internal dynamics). When the MSESS is applied to morphological predictions, the normalization is achieved by the mean-squared bed changes from the reference prediction, often the initial bed, which must adequately neutralize the inherent ease or difficulty of the prediction situation. In the remainder of this section we will challenge the efficacy of the normalization.

Let us consider a natural morphological development in time with a mean-squared difference MSE_r with the start situation that increases linearly in time. A prediction of this development obtains a constant skill throughout the simulation if the MSE of the model is a constant fraction of the MSE_r (left panel of Fig. 3.5). In other words, larger errors are allowed for larger cumulative (net) change, probably in agreement with a modeller's first intuition that the difficulty of the prediction increases with the larger changes associated with longer lead times. In this way, the cumulative nature of morphology as the persisted variable in combination with the zero change model provides a built-in, progressive lowering of the (metaphorical) bar. Below, we demonstrate using hypothetical examples that it is not generally fair that the cumulative, net change from the reference determines the zero point at the scale of skill.

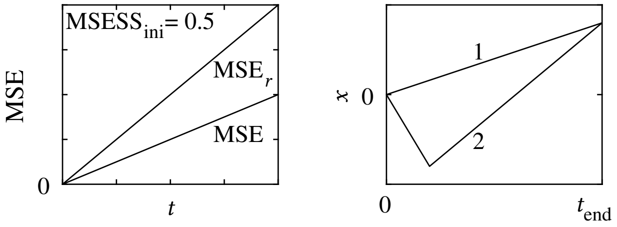


Figure 3.5: Effect of normalization by net bed changes. *Left*: for a steady increase of MSE_r in time, a constant skill of 0.5 is obtained when the MSE increases at half the speed of the MSE_r. *Right*: two pathways x of morphological features with a different prediction difficulty. The net changes from $t = 0$ to $t = t_\text{end}$ are equal and cannot discern between the two.

First, consider case P5 described in Sect. 3.3.5, which has a nonzero sediment budget error γ' due to a change of the mean of the prediction that is not present in the observations. Upon reversal of o' and p' (case R1), the sediment budget error also counts towards MSE_r (since now $\epsilon' = \gamma'$). As a consequence, a higher skill score was obtained, showing the lack of symmetry of the MSESS_ini. Apparently, there is a reward (in terms of a higher skill score) if the observations show a mean trend.

Next, we refer to and illustrate two hypothetical examples first described in Bosboom et al. (2014), i.e. Ch. 2. The first example concerns two different morphological developments, illustrated in the right panel of Fig. 3.5. Either development consists of a feature that has propagated to the same final position, such that also the net displaced sediment volumes are identical. Feature 1 has propagated at a steady speed to its final position, whereas feature 2 has first moved in the opposite direction under the influence of an episodic event, and, subsequently, moved back, under milder conditions. The latter is most likely considered a more difficulty prediction situation by a modeller. Nonetheless, cumulative (net) changes from the reference cannot discern between the two situations: the MSE$_r$ is the same.

The second example is of a summer–winter profile cycle over the course of five years, with small, random variations between the same seasons in consecutive years (Fig. 3.6, left panel). Assume that a model aimed at mimicking this behaviour is initialized from a winter profile, such that MSE$_r$ is smaller in the winter seasons than in summer seasons. For constant MSE, the diagnosed temporal evolution of model skill (Fig. 3.6, right panel) will demonstrate an artificial seasonal cycle with higher skill in summer, but with small changes between the same seasons over time from year to year. Clearly, a higher accuracy is required to obtain a certain skill level if the net observed changes are small. This may explain the low skill scores in van Rijn et al. (2003) for the seasonal morphology at Egmond for periods in which the breaker bar is relatively stable.

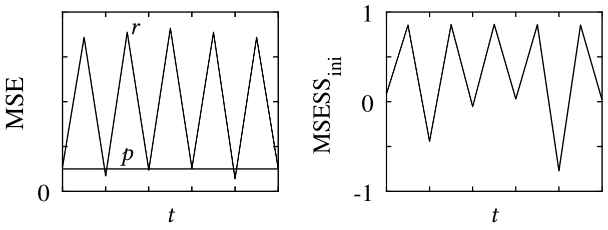


Figure 3.6: Skill for a hypothetical seasonal system. *Left*: the MSE for the prediction p is constant in time and the MSE$_r$ of the zero change reference prediction r fluctuates. *Right*: the MSESS$_{\mathrm{ini}}$ also fluctuates.

In both examples, the net change from the initial bed is not a proper indicator of the difficulty of the predictions; it lacks information on the nature of the morphological development prior to the evaluation time. For seasonal systems, an alternative choice of reference, for example the initial state for the same season, could have resulted in a more appropriate skill trend. However, even then, the progressive development is put to a less stringent test than the seasonal system, and increasingly so in time; the increase of MSE$_r$ with time provides the progressive development with an unfair advantage over the seasonal system for which the variation of MSE$_r$ is bounded, regardless of the amount of absolute change. In real-life situations, the behaviour of MSE$_r$ will be affected by a combination of trends and fluctuations. For instance, Walstra et al. (2012) presented predictions of cyclic net offshore bar migration that increase in skill from 0.38 after 1 year to 0.65 after 3 years as a consequence of an increase of MSE$_r$.

3.6.3 Large, persistent scales versus smaller scales

Several authors have reported skill scores of process-based morphodynamic models to increase with time. For a hindcast of the Western Scheldt from 1860 to 1970, skill scores steadily increase during the simulation from negative scores after about 20 years to $\text{MSESS}_{\text{ini}} \approx 0.5$ in 1970 (Dam et al., 2013, 2015). A hindcast of the morphological response of the Sand Engine during the first year of its existence showed negative values in the first two months ($\text{MSESS}_{\text{ini}} = -2$ in September and -0.03 in October, respectively), partly attributed to interpolation errors of the initial bathymetry, increasing to $\text{MSESS}_{\text{ini}} = 0.4$ in January and 0.59 in August (Luijendijk et al., 2017). A similar time-variation of skill was noticed for the 15 year hindcast of the evolution of the Bornrif, already mentioned in Sect. 3.6.1. Upon comparison of the Bornrif's yearly bed changes with the cumulative bed changes from the start of the simulation, Bosboom et al. (2014), see Ch. 2, noticed that the simulations capture little of the year-to-year variability while the larger-scale fields of cumulative change are reasonably well predicted. This was confirmed by an alternative skill computation that resulted in significantly lower skill values by considering bed changes in a one-year period rather than cumulative change over multiple years. Based on these findings, it was hypothesized that the relatively low values of $\text{MSESS}_{\text{ini}}$ at the beginning of the Bornrif simulation are mainly due to unskilful smaller scales. Over time, the relative contribution of these smaller scales to the cumulative change, and thus to $\text{MSESS}_{\text{ini}}$, decreases and, consequently, the contribution of the more skilful, persistent larger-scale trend increases.

To demonstrate this effect, a simple example is used. Assume the observed and predicted anomalies to consist of two spatial scales $o' = o'_{\text{high}} + o'_{\text{low}}$ and $p' = p'_{\text{high}} + p'_{\text{low}}$, respectively, with the subscripts *high* and *low* referring to a small and a large scale respectively. On the smaller scale, the anomalies are modelled by a sinusoidal variation in the entire model domain with $\rho_{p'o'} = -0.28$ and $\sigma_{p'}/\sigma_{o'} = 0.5$. Systematic biases are neglected, such that the skill for the small scale only is $\text{MSESS}_{\text{ini}} = -0.53$ throughout the simulation. For the larger scale, we have chosen a localized sedimentation and mirrored erosion feature, which steadily develop in size throughout the simulation, with $\rho_{p'o'} = 0.71$ and $\sigma_{p'}/\sigma_{o'} = 0.5$ leading to a constant skill of $\text{MSESS}_{\text{ini}} = 0.5$. The combined signals o' and p' are shown in Fig. 3.7 at the beginning, half-way and at the end of the simulation. Fig. 3.8 shows that the skill of the combined signal increases with time from $\text{MSESS}_{\text{ini}} = -0.08$ to 0.46. Hence, even when the skill of both the smaller and longer scales alone is constant with time, the combined skill increases from low scores at the beginning, dictated by the unskilful small scales, to higher scores towards the end, dictated by the more skilful longer scales.

In conclusion, an increase in skill for longer prediction horizons may well be indicative of the emerging of the more skilful larger scales, without the skill on these scales necessarily increasing in time. By implication, larger skill scores could be

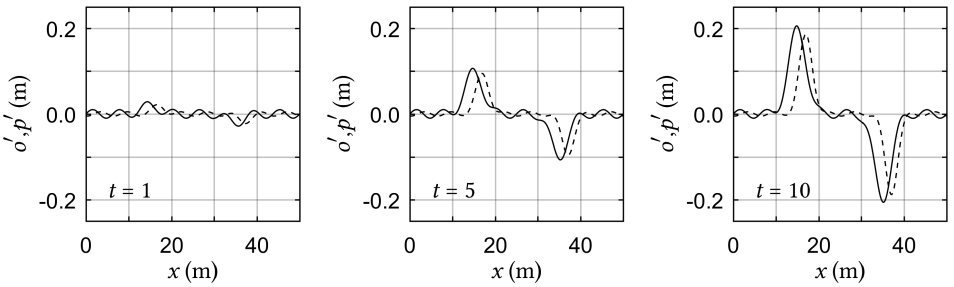


Figure 3.7: Measured (solid lines) and predicted anomalies at three moments (t = 1, 5 and 10) in the simulation, consisting of a constant high-frequent part and a low-frequent part that increases in magnitude in time.

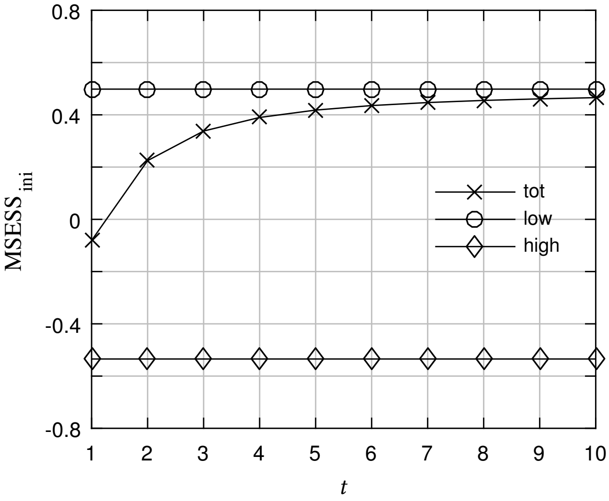


Figure 3.8: Variation of $\mathrm{MSESS}_{\mathrm{ini}}$ in time for the low-frequent (*low*) and high-frequent (*high*) part as well as for the combined signal (*tot*).

achieved from the beginning of the simulation by low-pass filtering of the sedimentation/erosion fields. The above not only challenges the comparability of skill scores at different times in a simulation but between different simulations as well. Clearly, there is a need to develop validation tools that distinguish between various spatial scales (see e.g. Bosboom and Reniers, 2014a, i.e. Ch. 6).

3.7 The BSS and beyond

As an easy-to-compute, normalized summary statistic, the BSS, or more appropriately MSESS, has become widely accepted for classification of morphodynamic model quality. A score of 0.5 means that the prediction is half-way an unskilful reference prediction and a perfect prediction, as measured by the cumulative mean-squared bed difference. Variations to the MSESS exist that aim to correct for measurement error, but these are inconsistent in either their skill formulation or their suggested classification scheme (Sect. 3.5.2).

However attractive the concept of skill, a comparative analysis based on skill scores lacks general validity. This is because cumulative bed change cannot adequately discern between the inherent ease or difficulty of predictions that have a different morphological development prior to the evaluation time—for instance trend, cyclic, episodic or different seasons in a seasonal system (Sect. 3.6.2). Also, due to the cumulative nature of morphology as the persisted variable, skill values are affected by an increasing dominance of persistent larger scales as simulations progress (Sect. 3.6.3). Hence, skill scores must be judged on a case-by-case basis.

When model calibration is aimed at maximizing the MSE skill score, the variance of the bed changes tends to be underestimated (Sect. 3.6.1). This is the direct consequence of the choice for a point-wise accuracy measure. It is advised to not only consider MSE and the MSE skill score, but also the error terms of the Murphy– Epstein decomposition (Murphy and Epstein, 1989), the anomaly correlation and ratio of the standard deviation of the predicted and observed anomalies. With this additional information, a more informed choice is possible for the most appropriate prediction given the goal at hand.

More in general, by using a single or a few accuracy or skill measures, only certain aspects of model quality are emphasized. In order to capture the various dimensions of morphodynamic model quality, multiple performance measures must be used. Welcome additions to existing performance measures would be methods that selectively address multiple spatial scales as well quantify the agreement in patterns and features rather than the point-wise agreement (see Bosboom and Reniers, 2014b, i.e. Ch. 4, and references therein). This requires an investment from the modelling community in the development of a more extensive model validation suite. The introduction of the BSS was a first step showing the maturing of the field of morphodynamic modelling (Roelvink and Reniers, 2012). The development of a set of performance measures in combination with a set of internationally agreed validation cases, as also advocated in Mosselman and Le (2016), would be an important next step in raising the level of morphodynamic model validation.

4 Displacement-based error metrics for morphodynamic models

This chapter is republished with minor changes only from J. Bosboom and A.J.H.M. Reniers (2014). Displacement-based error metrics for morphodynamic models. *Advances in Geosciences 39*, 37–43, doi:10.5194/adgeo-39-37-2014.

Chapter 2 (Bosboom et al., 2014) and Ch. 3 (Bosboom and Reniers, 2018) identified the need for performance measures that quantify the agreement in patterns and features rather than the point-wise agreement as well as the need for performance measures that selectively address multiple spatial scales. While the latter is addressed in Ch. 6 (Bosboom and Reniers, 2014a), the first point has led us to develop various error metrics that take the spatial structure of 2D morphological fields into account through a transformation of the computed towards the observed field (Ch. 4, this chapter, and Ch. 5). The pattern matching in this chapter optimizes the location of pixels with given intensities (i.e. depth values) in an image and is therefore probably closest to the visual validation by morphologists. By implication, the method is not sediment-conserving, as opposed to the optimization method of Ch. 5 (Bosboom et al., 2019), which moves sediment rather than depth values. The highlights of Ch. 4 are:

1. A novel diagnostic tool for the spatial validation of morphological fields is presented.
2. The method deforms the predictions such as to minimize the misfit with observations.
3. Errors are formulated based on the smooth displacement field between predictions and observations and on the residual point-wise error field.
4. Two new error metrics are introduced: a mean location error $\overline{D}$ and the $\mathrm{RMSE_w}$, which combines location and intensity errors.
5. The tool is tested against Delft3D model outcomes of the development of a idealized tidal inlet.
6. The new validation approach outperforms the convential approach based on the point-wise root-mean-squared error (RMSE)

Abstract

The accuracy of morphological predictions is generally measured by an overall point-wise metric, such as the mean-squared difference between pairs of predicted and observed bed levels. Unfortunately, point-wise accuracy metrics tend to favour featureless predictions over predictions whose features are (slightly) misplaced. From the perspective of a coastal morphologist, this may lead to wrong de-

cisions as to which of two predictions is better. In order to overcome this inherent limitation of point-wise metrics, we propose a new diagnostic tool for 2D morphological predictions, which explicitly takes (dis)agreement in spatial patterns into account. Our approach is to formulate errors based on a smooth displacement field between predictions and observations that minimizes the point-wise error. We illustrate the advantages of this approach using a variety of morphological fields, generated with Delft3D, for an idealized case of a tidal inlet developing from an initially highly schematized geometry. The quantification of model performance by the new diagnostic tool is found to better reflect the qualitative judgement of experts than traditional point-wise metrics do.

4.1 Introduction

Quantitative validation methods for morphodynamic models are often grid-point based; they compare observations and predictions per grid point and compute various metrics for the entire set or subset of grid points (e.g. Sutherland et al., 2004). Unfortunately, point-wise accuracy metrics, such as the commonly used MSE (mean-squared error) and RMSE (root-mean-squared error), tend to penalize, rather than reward, the model's capability to provide information on features of interest, such as scour holes, accumulation zones and migrating tidal channels. For instance, a prediction of a morphological feature that is correct in terms of timing and size, but is misplaced in space, may not outperform even a flat bed, which is inconsistent with the common judgement of morphologists (Fig. 4.1). This "double penalty effect" (Bougeault, 2003), which applies in full when a feature is misplaced over a distance equal or larger than its size, makes it difficult to demonstrate the quality of a high-variability prediction (Anthes, 1983; Mass et al., 2002). Clearly, a high-quality validation process requires alternative validation techniques that account for the spatial structure of 2D morphological fields.

For the verification of weather variables (e.g. precipitation), methods are being actively developed to quantify forecast performance based on spatial structure; see for instance Casati et al. (2008) and Gilleland et al. (2009) for an overview. One of the approaches in meteorology, now also pioneered in other fields (e.g. Haben et al., 2014; Ziegeler et al., 2012), is to find an optimal deformation of the predictions that minimizes the misfit with observations. This optimal deformation can be obtained by employing one of many existing image matching methods, of which optical flow techniques, designed to estimate motion, are probably most well-known in the coastal community. The result of the image matching or warping is a vector field of displacements, which can be regarded as a displacement error field. In addition, an intensity or amplitude error field may be defined as the difference between the deformed prediction and the observations (e.g. Marzban

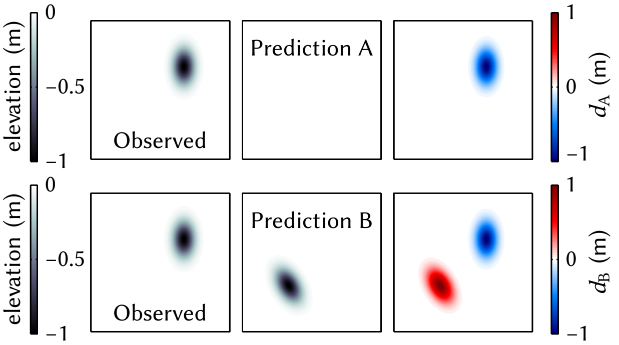


Figure 4.1: The “double penalty effect”. *Top panels*: the featureless prediction A has a nonzero difference d_A between predicted and observed depth values at the location of the observed feature only. *Lower panels*: prediction B, which reproduces the feature at the wrong location, is penalized twice (d_B is nonzero both where the predicted feature is and where it should be) and is thus diagnosed with a twice as large MSE[1].

and Sandgathe, 2010), which can be seen as the point-wise error if no penalty applies for misplacements.

Existing verification methods, based on field deformation of meteorological fields, not only differ in the applied image matching method, but also in the approach to the subsequent extraction of map-mean errors. Keil and Craig (2009) determine RMS (root-mean-squared) intensity and mean displacement errors within the boundaries of precipitation features, which they then combine into a single error metric. The latter requires the normalization of the two errors to put each term on equal footing, which introduces two parameters to the formulation. In contrast, Gilleland et al. (2010b) propose a combined error metric that besides the post-warp RMS intensity error and the mean displacement error also takes the original RMS intensity error into account, enabling a more fair comparison of forecast performance. Their metric, however, is not easily applicable since it requires three user-chosen weights that are dependent on the error terms themselves.

The goal of this paper is to quantify morphodynamic model performance, while taking the spatial characteristics of 2D morphology into account. Using a field deformation technique, we have developed and tested a new diagnostic tool for the validation of 2D morphological predictions. It includes a location (displacement) error metric and a robust and physically intuitive combined error metric that incorporates both location and intensity error. The combined metric rewards predictions to the degree that a larger error reduction can be obtained with smaller displacements. As a reference, we use the subjective but very powerful method of visual inspection of morphological patterns by experts.

[1] Corrected from Bosboom and Reniers (2014b) where (R)MSE was written.

Our method is outlined in Sect. 4.2, along with a brief description of the image warping method that we have adopted to calculate the optimal deformation. Next, in Sect. 4.3, we put the new diagnostic tool to the test, using morphological fields generated with Delft3D for an idealized case of a tidal inlet developing from an initially highly schematized geometry. Section 4.4 concludes with a summary of our findings and the implications for morphodynamic model validation.

4.2 Method

This section outlines our two-step approach in order to quantify the (dis)agreement between 2D morphological patterns. Section 4.2.1 describes the first step of deforming (or warping) the predicted morphology to minimize the point-wise error with observations. Next, Sect. 4.2.2 formulates two new error metrics, a mean location error that is distilled from the displacement vector fields and a single-number error metric that measures both the correspondence with respect to location and intensity (i.e. depth values).

4.2.1 Warping method

The measure of closeness between images or spatial fields is encountered in many fields from radiography to meteorology. This has led to the development of a multitude of image matching methods that, depending on the scientific field, are also named registration or warping methods. The goal of such methods is to find the optimal transformation that maps each point of a static image to a corresponding point (with the same intensity) in the moving image. Within the context of morphodynamic model validation, the static image represents the observed depth field o and the moving image the predicted depth field p.

Of all the available techniques, the class of optical flow techniques, designed to estimate small displacements in temporal image sequences, is probably the most well-known in our field. The basic assumption of optical flow is that the intensity of a moving object does not change appreciably in the considered time interval. We employ the efficient, nonrigid (i.e. allowing for free-form deformations) registration technique named Demon's registration (Thirion, 1998), which bears similarities to optical flow, in an implementation by Kroon and Slump (2009). The Demon's approach can be considered as similar to a minimization of the sum of square image intensities between the deformed predictions and observations (Pennec et al., 1999). It is therefore consistent with our quest to find the optimal deformation of the predictions that minimizes the point-wise (R)MSE.

The estimated backward pixel displacements $\mathbf{B}^* = (B^*_x, B^*_y)$ that are required for a given point in a static image (the observations in our validation context) to match the corresponding point in a moving image (the predictions) are given by

Thirion (1998):

$$\mathbf{B}^{*} = \frac{\left(I_p - I_o\right)\nabla I_o}{\left|\nabla I_o\right|^2 + \alpha^2\left(I_p - I_o\right)^2} \tag{4.1}$$

in which α is a normalization factor that is equal to 1 in the original method and I_o and I_p are the intensities of the static and moving image, respectively. The latter are taken as the observed and predicted depth fields, normalized by scaling between 0 and 1. Since Eq. 4.1 is based on local information, it is solved iteratively while including Gaussian smoothing as a regularization criterion. This ensures that a realistic, smooth displacement field is found instead of an irregular field that nonetheless minimizes the sum of squares. The normalization factor is chosen as $\alpha = 2.5$ in line with Kroon and Slump (2009) and the standard deviation of the Gaussian smoothing window as $\sigma = 4$. These parameters are kept constant for all registrations presented in Sect. 4.3. The forward displacements $\mathbf{F}^{*} = (F_x^{*}, F_y^{*})$ from the moving to the static image can be determined from $\mathbf{B}^{*}$ after the registration. Note that when in the following the subscript $*$ is dropped, we refer to the displacement fields transferred to a physical distance.

For the purpose of model validation, we interpret $d_0 = p_0 - o$, with p_0 the prediction prior to warp, as the total point-wise error and $d_1 = p_1 - o$, with p_1 the deformed prediction as follows from the registration, as the point-wise error if no penalty is imposed for location disagreement. Next, we use this perspective in the formulation of map-mean errors.

4.2.2 Formulation of new error metrics

From the Demon's registration (see Sect. 4.2.1), we obtain the optimal displacement vector field between predictions and observations as well as the optimal deformation of the predictions. "Optimal" in this context means that the sum of squares between the deformed predictions and observations is minimized, such that $0 \leqslant \text{RMSE}_1 \leqslant \text{RMSE}_0$, where RMSE_0 and RMSE_1 are the root-mean-squared errors before and after the warp, respectively. Note that we have preferred the RMSE over the MSE, since the first is measured in the same units as the data. Out of two predictions that have the same RMSE_0, a prediction that has similar morphological features as the measurements, albeit displaced, may receive a lower RMSE_1 than a prediction that is not able to reproduce the observed morphological features at all. Thus, the RMSE_1 is expected to diagnose the agreement between morphological fields if a zero penalty is imposed for misplacements of features. However, which of the two predictions is valued the better prediction by morphologists not only depends on RMSE_0 and RMSE_1, but also on the magnitude of the displacements required to obtain the error reduction. Therefore, we expect that the similarity in both location and intensity between morphological patterns can be fully assessed using three error metrics in concert: RMSE_0, RMSE_1 and a mean location error $\overline{D}$ that we will formulate next from the displacement vector fields.

It is tempting to define $\overline{D}$ as the arithmetic mean of $D = \sqrt{(B_x{}^2 + B_y{}^2)}$, the field of displacement magnitudes. However, it should be realized that the optical flow problem is underconstrained; for a single grid point, we only have information on the displacements normal to the contour lines, whereas along the contour lines the displacements are ambiguous (the so-called aperture problem). In the Demon's approach, the Gaussian smoothing acts as the necessary additional constraint, requiring that nearby grid points have similar displacements. As a consequence, nonzero displacements may be found along depth contours in morphologically inactive regions (see Sect. 4.3), whereas these displacements do not improve the match between the deformed prediction and the observations. Therefore, we propose a weighted mean location error that weights the local backward displacement magnitudes D with their effect on the reduction of the local squared error. In this way, displacements are only taken into account to the extent that they contribute to the minimization of the sum of squares. This yields:

$$\overline{D} = \frac{\sum_{i=1}^{n} w_i D_i}{\sum_{i=1}^{n} w_i}; \quad w_i = \frac{\mathrm{SE}_{0,i} - \mathrm{SE}_{1,i}}{\sum_{i=1}^{n}\left(\mathrm{SE}_{0,i} - \mathrm{SE}_{1,i}\right)}. \tag{4.2}$$

Here $\mathrm{SE}_0 = (p_0 - o)^2$ and $\mathrm{SE}_1 = (p_1 - o)^2$ are the local squared errors before and after the warp, respectively, n is the number of equidistant points in the spatial domain and $\sum_{i=1}^{n} w_i = 1$. Note that $\mathrm{RMSE}_j = \sqrt{n^{-1} \sum_{i=1}^{n} \mathrm{SE}_{j,i}}$, with $j = [0, 1]$.

Whereas model performance is usually diagnosed based on RMSE_0 only, we now have two additional metrics RMSE_1 and $\overline{D}$. In Sect. 4.3, it is demonstrated that considering these three metrics in concert allows a full assessment of model quality, avoiding the double penalty effect for misplaced features. In practice, guidance may be required on how to weight these three metrics. Besides, the morphologist may sometimes desire a single-number summary of model performance, especially if automated calibration routines are used. To serve these needs, we propose an adjusted RMS error measure, RMSE_w, that is computed from a field of weighted squared errors SE_w. The latter are determined by locally weighting SE_0 and SE_1. The purpose of the weighting procedure is to locally relax the requirement of an exact match to an extent determined by the local displacement magnitude. Figure 4.2 illuminates the weighting procedure for the ith grid point; an error reduction is awarded that is a fraction $1 - \delta_i$ of the full error reduction potential $(\mathrm{SE}_{0,i} - \mathrm{SE}_{1,i})$. Here, $\delta_i = D_i / D_\mathrm{max}$ and D_max is a maximum displacement length above which no relaxation is allowed. A larger fraction $1 - \delta_i$ is allowed for smaller displacement magnitudes D_i, with a maximum of $1 - \delta_i = 1$ and thus $\mathrm{SE}_{\mathrm{w},i} = \mathrm{SE}_{1,i}$ for $D_i = 0$ m. For $D_i \geqslant D_\mathrm{max}$, we have $1 - \delta_i = 0$ and thus $\mathrm{SE}_{\mathrm{w},i} = \mathrm{SE}_{0,i}$. Note that D_max is a user-defined, physically intuitive parameter that is dependent on the prediction situation and the goal of the simulation. It can be seen as the maximum distance over which morphological features may be displaced for the prediction to still get (some) credit for predicting these features. We now have for

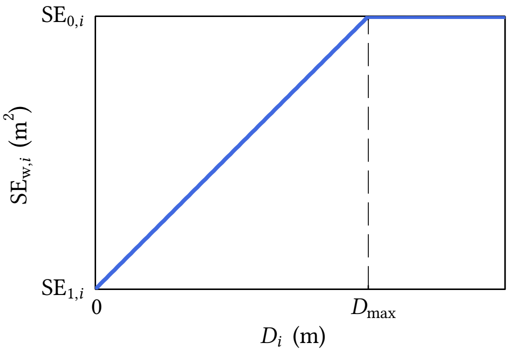


Figure 4.2: Weighted squared error for the ith grid point $\mathrm{SE}_{\mathrm{w},i}$, which is the sum of the local squared error after the warp $\mathrm{SE}_{1,i}$ and a penalty for misplacements $\delta(\mathrm{SE}_{0,i} - \mathrm{SE}_{1,i})$ with $\delta = D_i/D_{\max}$. The penalty ranges from 0 for $D_i \longrightarrow 0$ to $(\mathrm{SE}_{0,i} - \mathrm{SE}_{1,i})$ for $D_i = D_{\max}$, a user-defined maximum displacement length. For $D_i \geqslant D_{\max}$ the full point-wise error applies and $\mathrm{SE}_{\mathrm{w},i} = \mathrm{SE}_{0,i}$.

$\mathrm{RMSE}_{\mathrm{w}}$:

$$\mathrm{RMSE}_{\mathrm{w}} = \sqrt{\frac{\sum_{i=1}^{n} \mathrm{SE}_{\mathrm{w},i}}{n}} \tag{4.3}$$

where

$$\mathrm{SE}_{\mathrm{w}} = \mathrm{SE}_1 + \delta\big(\mathrm{SE}_0 - \mathrm{SE}_1\big) \tag{4.4}$$

$$\delta_i = \frac{D_i}{D_{\max}} \;\text{ for }\; D_i \leqslant D_{\max};\;\; \delta_i = 1 \;\text{ for }\; D_i > D_{\max}. \tag{4.5}$$

In conclusion, $\mathrm{RMSE}_{\mathrm{w}}$ as an error metric rewards forecasts to the degree that a larger error reduction can be obtained by smaller displacements. By definition, $\mathrm{RMSE}_1 \leqslant \mathrm{RMSE}_{\mathrm{w}} \leqslant \mathrm{RMSE}_0$. If the error reduction due to the image deformation is negligible or can only be obtained with displacements equal to or larger than $D_{\max}$, the diagnosed error is equal to the original error prior to the deformation RMSE_0. If, on the other hand, the displacements required to minimize the point-wise error are very small relative to $D_{\max}$, we have $\mathrm{RMSE}_{\mathrm{w}} \approx \mathrm{RMSE}_1$. The justification for this approach lies in the tendency of coastal morphologists to credit a prediction for the reproduction of features, albeit displaced, while imposing a relatively small penalty for misplacement. The intuitive weighting of these two aspects is mimicked by the user-defined parameter $D_{\max}$.

4.3 Application

Below, the new error metrics are used to diagnose the correspondence between model-generated pairs of morphological patterns for an idealized tidal inlet as well as the relative ranking between the pairs. The fields have been generated for the idealized case of a tidal inlet developing from an initially highly schematized geometry (Roelvink, 2006). First, Sect. 4.3.1 demonstrates that the location error $\overline{D}$ is able to capture the overall misplacement of the morphological patterns. Next, in Sect. 4.3.2, the combined error metric $\mathrm{RMSE_w}$ is put to the test. Two examples are shown where the $\mathrm{RMSE_w}$ makes the right the decision as to which of two predictions is the better prediction while the conventional, purely point-wise $\mathrm{RMSE_0}$ fails to do so.

4.3.1 Location error

In this subsection, we consider a subset of the model-generated depth fields which only differ with respect to the latitude, and hence Coriolis parameter, used in the model. Of four depth fields, we label the field generated at 53° N as the "observations" (Fig. 4.3a) and consider the other fields, for latitudes 90° N, 0° and 90° S, as three competing predictions. Even though the predictions are not shown here, it will not come as a surprise that the point-wise error $\mathrm{RMSE_0}$ is smallest for 90° N and largest for 90° S (Table 4.1).

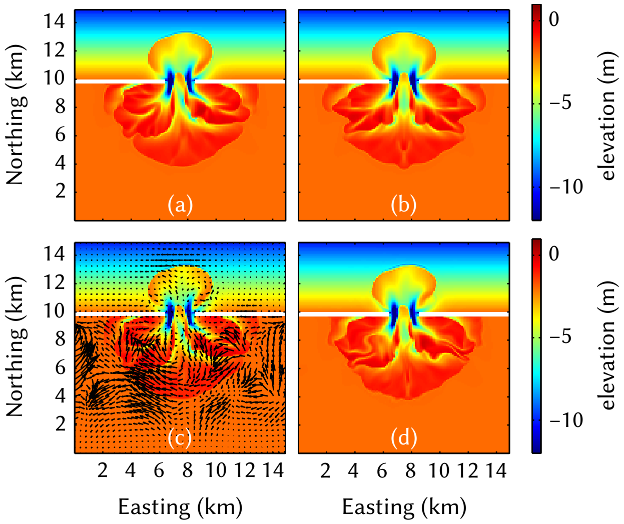


Figure 4.3: Example of the image warp: **(a)** the "observations", calculated using Delft3D with Coriolis at 53° N, **(b)** the predictions, calculated at 0°, **(c)** the backward displacement vector field $\mathbf{B}$ of the observations towards the predictions, shown on top of the observations, and **(d)** the predictions deformed to more closely match the observations.

In order to determine $\mathrm{RMSE_1}$ and $\overline{D}$, the image warping method is applied, following the procedure outlined in Sect. 4.2, and illustrated here for the prediction at 0° (Fig. 4.3b). The deformed prediction that matches the observations most closely is shown in Fig. 4.3d and the corresponding backward vector displacement field $\mathbf{B}$ in Fig. 4.3c. As explained in Sect. 4.2.1, in the inactive outer regions, physic-

Latitude	RMSE$_0$ (m)	RMSE$_1$ (m)	$\overline{D}$ (m)
90° N	0.29	0.12	180
0°	0.52	0.26	350
90° S	0.73	0.35	710

Table 4.1: Errors for competing predictions that differ with respect to the latitude, and thus the Coriolis parameter, used in Delft3D. The model results for 53° N are regarded as the "observations".

ally unrealistic displacements are found along depth contours, since no penalty is imposed in the minimization for displacements along depth contours. As will be illustrated next, this is solved for in the formulation of $\overline{D}$ (Eq. 4.2).

The difference d_0 between the predictions prior to the warp and the observations is shown in Fig. 4.4a, whereas Fig. 4.4b shows the difference d_1 after the warp. Note that taking the root-mean-square of d_0 and d_1 yields RMSE$_0$ and RMSE$_1$, respectively. From d_0, the double penalty problem is clearly observed; for instance at the edges of the ebb-tidal delta, an error is diagnosed both where the delta is present in the observations but absent from the predictions and vice versa. After the warp, both errors have practically disappeared, such that they will not count towards RMSE$_1$, demonstrating again that RMSE$_1$ should be regarded as the point-wise error if no penalty for misplacement is taken into account. For the prediction at 0°, RMSE$_1$/RMSE$_0$ = 0.5, and slightly smaller ratios are found for the other two predictions (Table 4.1).

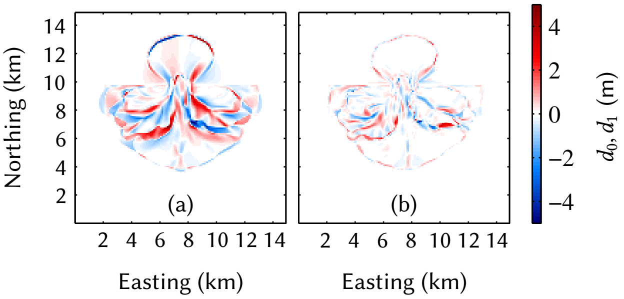


Figure 4.4: Point-wise error fields for the predicted depth field at 0°: **(a)** the total error $d_0 = p_0 - o$ before the warp, **(b)** the error $d_1 = p_1 - o$ after the warp, to be regarded as the remaining point-wise error if no penalty applies for location disagreement.

The weighted dispacements wD, with $D = \sqrt{(B_x{}^2 + B_y{}^2)}$ and w according to Eq. 4.2, are shown in Fig. 4.5. Inherent to the use of the squared error to determine w is that larger error reductions are heavily weighted. Here, we have nevertheless chosen this weighting since squared errors are consistent with the minimization as performed by the registration method as well as with the use of the (R)MSE as the point-wise metric, which is common in morphodynamic model validation.

Note that for the computation of $\overline{D}$ (Eq. 4.2), we require the backward (from the observations to the predictions) rather than the forward displacements; for each point in the observational domain, these provide the distance at which the point in the predictions is located that is shifted to the considered location in the observations. Summing wD for the entire domain yields a location error $\overline{D} = 350$ m at 0° (Table 4.1).

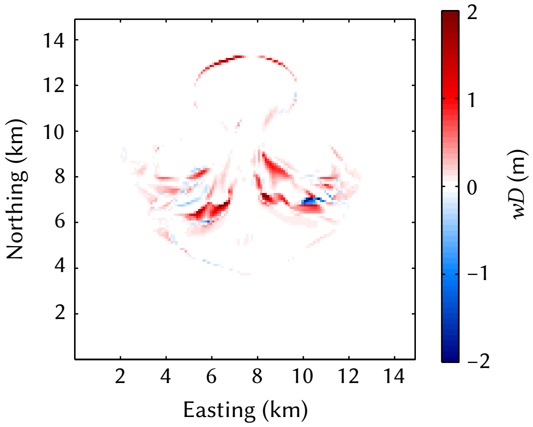


Figure 4.5: Weighted displacements wD for the prediction at 0°. Here $D = \sqrt{(B_x{}^2 + B_y{}^2)}$ is the field of displacement magnitudes computed from the backward displacement vector field $\mathbf{B}$ (see Eq. 4.1) and w is determined according to Eq. 4.2.

The values for $\overline{D}$ for the three predictions demonstrate a qualitative behaviour consistent with the error in latitude and hence Coriolis effect in the various predictions (Table 4.1). In fact, all three error metrics, RMSE_0, RMSE_1 and $\overline{D}$ diagnose the predictions for 90° N and 90° S as the best and worst predictions, respectively. Next, we will consider situations in which a ranking consistent with expert judgement is only obtained by considering these three metrics in concert, using an appropriate weighting, or from RMSE_w.

4.3.2 Ranking according to the combined error metric

In this subsection, we present an example, again using depth fields generated with the Delft3D model of the schematized tidal inlet, that demonstrates that RMSE_w outperforms the traditional score RMSE_0. Now, the model results at a latitude of 0° (see Sect. 4.3.1) are assumed to be the "truth". Four competing predictions are considered that are generated at 0° with various changes to the model boundary conditions (w.r.t. tidal amplitude and flow direction). Figure 4.6 shows the four predictions, the "observations" and the deformed predictions that minimize the point-wise error.

We have labelled the predictions according to a subjective ranking based on visual inspection, with A the prediction with the closest match with the observa-

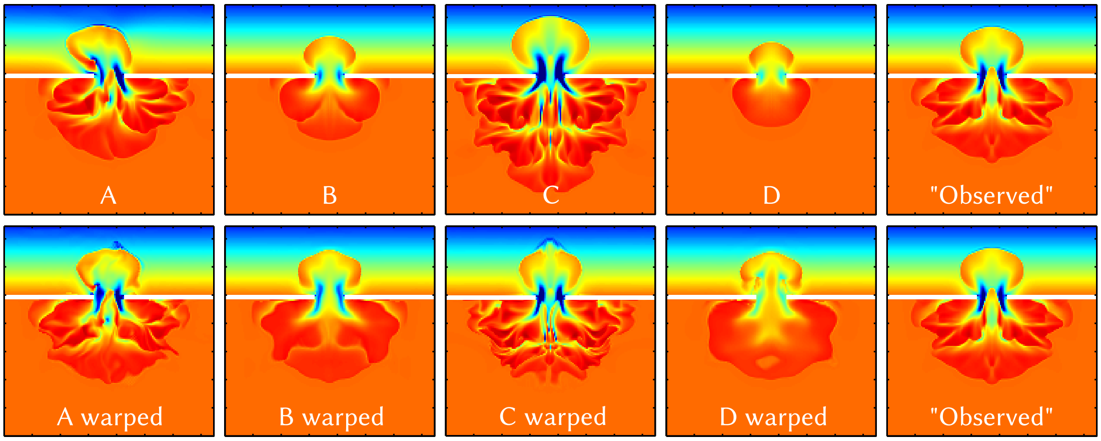


Figure 4.6: Predictions A, B, C and D, the "observations" (taken as the model results for 0°) and the corresponding deformed predictions that minimize the point-wise mismatch between predictions and observations. The labels are chosen such that the lower the label in the alphabet, the higher the quality that the prediction is probably diagnosed with upon visual inspection. The axes are as in Fig. 4.3.

tions and D, the worst prediction. We have a slight preference for prediction B over C, but it is possible that other morphologists would tend to regard C as the better prediction. Not surprisingly, the relative ranking as diagnosed by RMSE_0 deviates from the expert ranking (Table 4.2); based on RMSE_0 one would wrongfully conclude that predictions A and B perform equally well and that prediction D outperforms prediction C.

Prediction	Ranking	RMSE_0 (m)	RMSE_1 (m)	$\overline{D}$ (m)	RMSE_w (m)
A	1	0.78	0.38	610	0.49
B	2	0.77	0.53	770	0.60
C	3	1.16	0.56	860	0.78
D	4	0.95[2]	0.77	1230	0.84

Table 4.2: Subjective ranking (with 1 being the best prediction) and errors for competing predictions, generated with Delft3D for various boundary conditions. The "observations" are taken as the model outcome at 0° (cf. Sect. 4.3.1). The values for RMSE_w hold for $D_{\text{max}} = 3000$ m.

The values of RMSE_1, $\overline{D}$ and RMSE_w for the respective predictions provide the necessary additional information on model performance (Table 4.2). The smaller RMSE_1 for prediction A than for prediction B shows that if no penalty is imposed for misplacements, prediction A receives a lower error than B. Moreover, a smaller average displacement $\overline{D}$ is required to minimize the point-wise error. Thus, even though no distinction can be made based on RMSE_0, we can conclude that

[2] Corrected from Bosboom and Reniers (2014b) where 0.96 was listed.

pattern A more closely corresponds to the observations than pattern B. Clearly, considering the values of RMSE_0, RMSE_1 and $\overline{D}$ in concert leads to a diagnosis of relative model performance of A and B in line with visual inspection.

To determine RMSE_w, a value for D_max must be chosen. A defendable choice would be to limit D_max to the scale of the morphological features of interest. For this particular case, $D_\mathrm{max} = 3000\,\mathrm{m}$ is considered appropriate, being in the order of magnitude of the seaward extent of the ebb-tidal delta. In general, of course, D_max must be chosen in accordance with the goal of the simulation.

Figure 4.7 shows that with $D_\mathrm{max} = 3000\,\mathrm{m}$, RMSE_w reports a higher quality for prediction A than for prediction B, regardless of the exact choice for D_max. Only if one decides to not allow any relaxation of the requirement of an exact match ($D_\mathrm{max} = 0$ m), RMSE_w is identical to the full point-wise error RMSE_0 and no distinction can be made between A and B. If one wishes to allow the full error reduction potential ($D_\mathrm{max} \longrightarrow \infty$), we have $\mathrm{RMSE}_\mathrm{w} = \mathrm{RMSE}_1$.

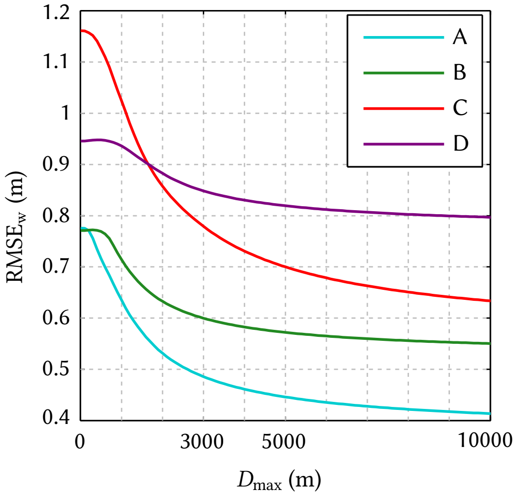


Figure 4.7: The combined error metric RMSE_w as a function of D_max for predictions A, B, C and D. The larger D_max, the more the requirement of an exact match is relaxed; for $D_\mathrm{max} \longrightarrow 0$, we have $\mathrm{RMSE}_\mathrm{w} \longrightarrow \mathrm{RMSE}_0$ and for $D_\mathrm{max} \longrightarrow \infty$, we have $\mathrm{RMSE}_\mathrm{w} \longrightarrow \mathrm{RMSE}_1$.

Table 4.2 illuminates that prediction C, the prediction with the largest RMSE_0, has a much larger potential for error reduction by warping than prediction D; notwithstanding the larger RMSE_0, RMSE_1 is smaller for prediction C than for D and at a smaller mean displacement $\overline{D}$. The relatively small error reduction potential for D is a result of the fact that features not present in the predictions remain absent after the warping procedure, as evident in the deformed predictions in Fig. 4.6. As a result, RMSE_1 remains relatively high for D, rightfully penalizing the prediction for the absence of the observed features. A conclusive answer as to whether C or D is the better prediction now requires a (subjective) weighting of RMSE_0, RMSE_1 and $\overline{D}$. Conveniently, the weighting between location errors, pre-warp and post-warp intensity errors is already provided by the formulation of $\mathrm{RMSE_w}$, allowing a quantitative single-number comparison between predictions C and D. For $D_{\max} = 3000\,\mathrm{m}$, the values for $\mathrm{RMSE_w}$ indicate that prediction C outperforms D (Fig. 4.7), consistent with the ranking based on visual inspection. Naturally, the occurence of this ranking reversal, as compared to the ranking based on $\mathrm{RMSE_0}$, depends on the chosen value of $D_{\max}$.

4.4 Conclusions

We have developed a new diagnostic tool for morphodynamic model validation. It employs an image warping method that finds the smooth displacement field between predictions and observations that minimizes the point-wise error. Two new metrics are proposed: (1) a location error $\overline{D}$ that is determined as a weighted mean distance between morphological fields; and (2) a combined error metric $\mathrm{RMSE_w}$ that takes both location and intensity errors into account.

A full appreciation of the quality of a prediction can be obtained when considering $\overline{D}$ in concert with both the original point-wise error $\mathrm{RMSE_0}$ and the point-wise error of the deformed predictions, $\mathrm{RMSE_1}$. In order to quantify the relative performance between predictions, a (subjective) weighting of these three metrics must be carried out. Alternatively, the weighting is already provided by $\mathrm{RMSE_w}$, which combines all relevant information on location errors and pre- and post-warp intensity errors.

The combined error metric credits predictions to the degree that a larger error reduction can be obtained with smaller displacements. It reduces to $\mathrm{RMSE_0}$ if all displacements are larger than a user-defined $D_{\max}$ and to $\mathrm{RMSE_1}$ for displacements that are negligible relative to $D_{\max}$. The latter can be seen as the maximum distance over which morphological features may be displaced for the prediction to still get (some) credit for predicting these features. The appropriate choice for $D_{\max}$ depends on the prediction situation and the goal of the simulation. Since it only requires a single, physically intuitive parameter, $\mathrm{RMSE_w}$ provides a robust basis for comparison.

An example of a schematized tidal inlet has demonstrated that $\mathrm{RMSE_w}$ outperforms the conventional validation approach based on a strictly point-wise metric such as $\mathrm{RMSE_0}$. In situations where morphological features are misplaced, point-wise accuracy metrics tend to favour predictions that underestimate variability. For the schematized tidal inlet, it was shown that, as opposed to $\mathrm{RMSE_0}$, the new combined error metric $\mathrm{RMSE_w}$ makes choices as to which of two predictions is better, which are consistent with visual validation by experts.

Acknowledgements

The authors wish to thank both reviewers for their constructive comments. Ian Townend (HR Wallingford[3]) is thanked for stimulating discussions and helpful comments.

[3] Now Visiting Professor at the University of Southampton.

5 Optimal sediment transport for morphodynamic model validation

This chapter has been submitted for publication as J. Bosboom, M. Mol, A.J.H.M. Reniers, M.J.F. Stive and C.F. de Valk (2019). Optimal sediment transport for morphodynamic model validation.

The combined error metric RMSE_w as presented in Ch. 4 (Bosboom and Reniers, 2014b) was seen to capture the visual closeness of morphological patterns, therewith highlighting aspects of model quality that are not reflected in point-wise metrics, such as the root-mean-squared error (RMSE). The underlying image warp minimizes the point-wise squared error by freely deforming the computations towards the observations. Since it essentially shifts individual depth values by (locally) stretching or compressing the morphological pattern, sediment mass continuity is not guaranteed. For morphodynamic model validation, it seems more natural to define the mismatch between predictions and observations in terms of the physical quantity responsible for morphodynamic development: sediment transport. This idea is at the heart of the here presented effective transport difference (ETD) method leading to a novel error metric, the root-mean-squared transport error (RMSTE). The ETD method, like the image warp, takes the spatial ordering of grid points and therewith the morphological pattern into account. It results in a perfect transformation of the predicted to the observed morphological field, whereas, due to bed level differences between corresponding features in the two fields, the image warp does not allow an exact match. As opposed to the warp, the ETD method is mass-conserving, parameter-free and symmetric, the optimal transport from observations to predictions being the inverse of the optimal transport from predictions to observations.

The highlights of this chapter are :

1. A novel diagnostic tool for morphodynamic model validation is presented that moves misplaced sediment from the predicted to the observed morphology through an optimal, rotation-free sediment transport field.
2. The optimal transport field is relatively easily found through a parameter-free procedure solving a Poisson equation.
3. A new error metric, the RMSTE, is defined as the root-mean-square of the optimal transport field.
4. The RMSTE, as opposed to the RMSE, is able to discriminate between predictions that differ in the misplacement distance of predicted morphological features.
5. The RMSTE avoids the consistent favouring of the underprediction of morphological variability that the RMSE is susceptible to.

Abstract

Although commonly used for the validation of morphological predictions, pointwise accuracy metrics, such as the root-mean-squared error (RMSE), are not well suited to demonstrate the quality of a high-variability prediction; in the presence of (often inevitable) location errors, the comparison of depth values per grid point tends to favour predictions that underestimate variability. In order to overcome this limitation, this paper presents a novel diagnostic tool that defines the distance between predicted and observed morphological fields in terms of an optimal sediment transport field, which moves the misplaced sediment from the predicted to the observed morphology. This optimal corrective transport field has the "cheapest" quadratic transportation cost and is relatively easily found through a parameter-free and symmetric procedure solving an elliptic partial differential equation. Our method, which we named effective transport difference (ETD), is a variation to a partial differential equation approach to the Monge–Kantorovich L^2 optimal transport problem. As a new error metric, we propose the root-mean-squared transport error (RMSTE) as the root-mean-squared value of the optimal transport field. We illustrate the advantages of the RMSTE for simple 1D and 2D cases as well as for more realistic morphological fields, generated with Delft3D, for an idealized case of a tidal inlet developing from an initially highly schematized geometry. The results show that by accounting for the spatial structure of morphological fields, the RMSTE, as opposed to the RMSE, is able to discriminate between predictions that differ in the misplacement distance of predicted morphological features, and avoids the consistent favouring of the underprediction of morphological variability that the RMSE is prone to.

5.1 Introduction

Quantitative validation methods for morphological predictions are often grid-point based: they compare observations and predictions per grid point and compute various metrics for the entire set or subset of grid points. Accuracy metrics, e.g. the root-mean-squared error (RMSE) or the mean absolute error (MAE) measure the averaged correspondence between individual pairs of model outcomes and observations, whereas skill metrics determine the accuracy, using an accuracy metric of choice, relative to the accuracy of a prediction produced by a standard of reference (Gallagher et al., 1998). Several morphological studies rely solely on a skill score, most notably a mean-squared-error skill score (MSESS or BSS, Sutherland et al., 2004; Bosboom et al., 2014, i.e. Ch. 2) as a performance metric (e.g. van Rijn et al., 2003; Plant et al., 2004; Henderson et al., 2004; Pedrozo-Acuña et al., 2006; Scott and Mason, 2007; Ruggiero et al., 2009; Orzech et al., 2011; Walstra et al., 2012; Williams et al., 2012; Simmons et al., 2017; Monge-Ganuzas et al., 2017; Luijendijk et al., 2017; Luijendijk et al., 2019). In other cross-shore, longshore and area modelling studies, skill scores and accuracy metrics are used in concert (e.g. Ruessink et al., 2007; Dam et al., 2016; Fortunato et al., 2014; Simmons et al., 2019). These procedures are sometimes supplemented with bias- and correlation-based measures, either directly (e.g. Gallagher et al., 1998; Roelvink et al., 2009; McCall et al., 2010, 2015; Ganju et al., 2011; Davidson et al., 2013; Dodet et al., 2019; Hallin et al., 2019) or through the Murphy-Epstein decomposition of the MSESS (Sutherland et al., 2004; Bosboom and Reniers, 2018, i.e. Ch. 3), which additionally employs an amplitude error (e.g. Sutherland et al., 2004; Ruessink and Kuriyama, 2008; van der Wegen et al., 2011; van der Wegen and Roelvink, 2012; Elmilady et al., 2019).

The various statistical measures condense a large number of data into a single value, inevitably emphasizing only certain aspects of the quality of the model results. Morphodynamic modellers are inclined to judge model results on the reproduction of patterns. Unfortunately, point-wise accuracy and derived skill metrics tend to penalize, rather than reward, the model's capability to provide information on features of interest, such as scour holes, accumulation zones and migrating bars or tidal channels (Bosboom et al., 2014; Bosboom and Reniers, 2018, i.e. Chs. 2 and 3, respectively). This tendency to reward the underestimation of variability (Anthes, 1983; Arpe et al., 1985; Taylor, 2001) is easily illustrated by the classical example of the "double penalty effect" (Bougeault, 2003): a prediction, which reproduces a feature at the wrong location, is penalized twice, both where the predicted feature is and where it should be, and is thus diagnosed with a twice as large mean-squared error (MSE) as a flat bed prediction. More in general, for a nonperfect correlation, as would be the case in the presence of location errors, accuracy as well as skill values can be "improved" by underestimation of the variability (Bosboom et al., 2014; Bosboom and Reniers, 2018, i.e. Chs. 2 and 3, respectively). Clearly, this is inconsistent with the common judgement of morphologists. In order to avoid the underestimation of bed changes, an indicator should be added to determine whether the predicted variance is close to the observed variance. Further, since point-wise metrics do not take the spatial ordering of grid points into account, they are not sensitive to misplacement distance. The simplest demonstration of the latter is a prediction of a feature on a otherwise flat bed that has been misplaced over a distance larger than its size. For this situation, metrics that impose a penalty on point-wise bed level differences yield identical values irrespective of the misplacement distance.

The above illustrates the need for new validation metrics that account for the spatial structure of morphological fields. Pioneering techniques in the field of weather forecasting comprise field deformation methods, which give information about how much the predicted field needs to be manipulated spatially (displacement, rotations, scaling, etc.) and quantify the residual errors (Gilleland et al., 2009). Bosboom and Reniers (2014b), i.e. Ch. 4, developed a field deformation or image warping approach for morphological model validation that determines a smooth displacement field between morphological predictions and observations minimizing the residual point-wise error and computes domain-averaged errors based on the displacement as well as residual error fields. The method includes a robust and physically intuitive combined error metric, the $\mathrm{RMSE_w}$, which rewards predictions to the degree that a larger error reduction can be obtained with smaller displacements. This error metric for morphological model validation results in choices as to which of two predictions is better that are consistent with visual validation, demonstrating the potential of field deformation methods to overcome the limitations of point-wise metrics. However, the so-determined optimal smooth transformation merely relocates predicted bed levels in the two-dimensional domain. As a result, horizontal dimensions of features may get distorted, such that sediment is not necessarily conserved.

For morphodynamic model validation, it seems more natural to base a validation metric on a transformation between predictions and observations defined in terms of the physical quantity responsible for morphodynamic development: sediment transport. Therefore, we have developed a method that determines the distance between morphological fields in terms of the minimal sediment transport required to change the one field into the other. Since the transformation is defined in terms of sediment transport, mass will now be conserved, but features may not. The optimal transformation or effective transport difference (ETD) has the “cheapest” transportation cost and is relatively easily found by solving an elliptic partial differential equation. The solution procedure is parameter-free and symmetric, the optimal transport field from observations to predictions being the inverse of the optimal transport field from predictions to observations. The new domain-averaged error metric that we propose is a multiple of the minimum transportation cost.

Our ETD method is related to the Monge–Kantorovich theory of optimal mass transport, which deals with the transport of a distribution of mass to another distribution of mass on the same space, in such a way as to keep the transportation cost to a minimum. The first formulation of the optimal mass transport problem was due to Monge in 1781, who considered the most economical way of transporting a pile of soil for construction works from one site to another. Monge used a cost function equal to the norm of the distance, based on the argument that the cost of transportation of an individual mass is proportional to its weight times the travelled distance (Rachev and Rüschendorf, 1998). This leads to the physical interpretation of the L^1 optimization, which minimizes the norm, in terms of the minimization of work, assuming that the work of transporting a mass element Δm over a distance Δd is $\Delta m \Delta d$ (Bogachev and Kolesnikov, 2012).

The work of Kantorovich in 1948 gave the optimal mass transport problem its modern, generalized formulation, which is today known as the L^p Monge–Kantorovich problem (Villani, 2003). Here one is allowed to “divide grains”, whereas in Monge’s formulation grains that share the same initial location must also share the same final location (Rachev and Rüschendorf, 1998). Especially the L^2 Monge–

Kantorovich problem, which minimizes the squared norm of the distance, has been researched intensively by theoretical mathematicians (Villani, 2003), since, as opposed to the L^1 problem, it allows for relatively simple solutions. The search for efficient numerical solvers has only recently become a lively research domain (Santambrogio, 2015). Benamou and Brenier (2000) and Benamou et al. (2002) were the first to construct a robust and efficient numerical solver for the L^2 Monge–Kantorovich problem by introducing a partial differential equation approach. In a fluid mechanics framework, they showed that the L^2 optimal transport is equivalent to minimizing a kinetic energy functional among solutions of the continuity equation, with the optimal solution given as the gradient of a potential and, thus, being irrotational.

Our ETD method also employs an irrotationality condition for the optimal transport in order to reformulate a transport optimization problem in terms of a partial differential equation that is easily solved. However, whereas the L^2 Monge–Kantorovich problem penalizes the quadratic distance the transformation moves each bit of material, weighted by the material's mass, our quadratic cost function penalizes the squared sediment transport, i.e. mass times distance, herewith retaining the original physical Monge's interpretation in terms of work, albeit in a quadratic sense. New aspects are further that our model boundaries are open to sediment, which allows a bias to exist between the two bathymetric fields.

This paper presents a novel error metric, the root-mean-squared transport error (RMSTE) as a multiple of the "cheapest" quadratic cost for the transportation of sediment from predictions to observations and establishes its applicability for morphodynamic model validation. In doing so, for fairness of comparison, the behaviour of the RMSTE as an error metric is evaluated in comparison to the behaviour of its point-wise counterpart, the RMSE. First, in Sect. 5.2, we describe our ETD method of finding the optimal transport difference between two morphological fields, leading to the formulation of the RMSTE. Section 5.3 shows in 1D how the RMSTE and RMSE behave for both misplaced features and features that are underestimated in size. Further, it compares the RMSTE and RMSE for a 2D example, in which the latter of these two metrics suffers from the double penalty effect. Next, in Sect. 5.4, we put the RMSTE to a more realistic test using morphological fields, generated with Delft3D, for an idealized case of a tidal inlet developing from an initially highly schematized geometry. This section not only compares the RMSTE to the RMSE but to the RMSE$_\mathrm{w}$ as well. The implications of the results for morphodynamic model validation are discussed in Sect. 5.5. Section 5.6 concludes with a summary of our findings and identifies future work.

5.2 A new method

In this section, we present a new error metric, the RMSTE, which measures the mismatch between two morphological fields in terms of sediment transport. First, Sect. 5.2.1 describes common error metrics, such as the RMSE, that penalize bed level differences between predictions and observations. Second, in Sect. 5.2.2, we define the RMSTE as (a multiple of) the optimal (i.e. minimum) quadratic transport cost required to transform the predictions into the observations. Third, Sect. 5.2.3 demonstrates that the optimal transport, on which RMSTE is based, can be found by solving an elliptic partial differential equation. Finally, Sect. 5.2.4 briefly describes the numerical implementation.

5.2.1 Penalty on bed level differences

More traditional error metrics are based on a point-wise comparison of predictions and observations. Let h_1 and h_2 be the predicted and observed bed levels above a certain vertical reference level, respectively, for a set of points $\mathbf{x}$ over a domain Ω. If $e = h_2 - h_1$ is the point-wise bathymetric error, the p-norm bathymetric error is defined as:

$$\|e\|_p = \left(\int_{\mathbf{x}\in\Omega} |e|^p \mathrm{d}\mathbf{x} \right)^{1/p} \tag{5.1}$$

with $p = 1, 2, \infty$ the usual choices for p and the $p = 2$ norm known as the *Euclidean norm*. Often used point-wise accuracy metrics, the MAE and the RMSE are constant multiples of the 1-norm and 2-norm errors, respectively. The RMSE reads:

$$\text{RMSE} = \frac{1}{\sqrt{A_\Omega}} \left(\int_{\mathbf{x}\in\Omega} |h_2(\mathbf{x}) - h_1(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x} \right)^{1/2} = \frac{1}{\sqrt{A_\Omega}} \|e\|_2 \tag{5.2}$$

with A_Ω the domain surface area. The MSE simply is the square of the RMSE.

Note that in Sects. 5.3 and 5.4, we have chosen to visually compare predictions and observations by means of difference fields $\delta = h_1 - h_2$ rather than error fields e. The advantage of defining the deviations as predicted values minus real, observed values is that the observations are the reference point from which the predictions may differ, such that a positive deviation indicates an overprediction and a negative deviation an underprediction. Of course, the RMSE is unaffected when computed from difference fields δ rather than from e.

5.2.2 Penalty on transport magnitude

Assume that $\mathbf{q}$ on Ω represents a cumulative, depth-integrated transport of sediment from h_1 to h_2, such that with a constant grain size and porosity, and, hence, constant density, the sediment volume balance is satisfied:

$$\nabla \cdot \mathbf{q} = h_1 - h_2 \tag{5.3}$$

with $\nabla\cdot$ is the divergence operator and either known or unknown transports normal to the boundary $\partial\Omega$ of Ω at every point of $\partial\Omega$. Note that $\mathbf{q}$ is to be interpreted as a corrective transport field moving sediment from the predicted morphology h_1 to the observed field h_2, and, hence, as a transport difference field between h_1 and h_2.

There may exist a multitude of transport fields satisfying Eq. 5.3. An *optimal* field can be determined by minimizing the p-norm of the transport field:

$$\underset{\mathbf{q}}{\text{minimize}} \quad \|q\|_p = \left(\int_{\mathbf{x}\in\Omega} |\mathbf{q}(\mathbf{x})|^p \,\mathrm{d}\mathbf{x}\right)^{1/p} \tag{5.4}$$

with $q = |\mathbf{q}|$ is the magnitude of the transport field.

Equation 5.4 under the constraint Eq. 5.3 differs from the L^p Monge–Kantorovich mass transfer problem (Villani, 2003) in that it minimizes the cumulative transport, and thus, assuming constant density, mass times distance, to the power p, rather than the travelled distance to the power p, weighted by the amount of transferred mass. If the exponent $p = 1$, the minimization problem of Eqs. 5.3 and 5.4 reduces to an L^1 Monge–Kantorovich problem with the Euclidean distance as the cost function (Evans, 1997). Numerical methods for solving this problem exist (Benamou and Carlier, 2015), but are considerably more complex than the solution of Eqs. 5.3 and 5.4 with $p = 2$. As we will see in Sect. 5.2.3, the case $p = 2$ is relatively easily solved by rewriting the optimality condition Eq. 5.4 and will therefore be the one used in this paper.

Summarizing, we will solve the following L^2 problem minimizing a quadratic transport cost:

$$\begin{aligned} &\underset{\mathbf{q}}{\text{minimize}} \quad \|q\|_2 = \left(\int_{\mathbf{x}\in\Omega} |\mathbf{q}(\mathbf{x})|^2 \,\mathrm{d}\mathbf{x}\right)^{1/2} \\ &\text{subject to} \quad \nabla \cdot \mathbf{q} = h_1 - h_2. \end{aligned} \tag{5.5}$$

By rewriting the cost functional, Eq. 5.5 can be reformulated as a an elliptic partial differential equation from which the quadratic optimal transport field $\mathbf{q}_{L2}$ is relatively easily solved (see Sect. 5.2.3). In analogy with Eq. 5.2, we can now introduce the RMSTE as:

$$\mathrm{RMSTE} = \left(\frac{1}{A_\Omega} \int_{\mathbf{x} \in \Omega} |\mathbf{q}_{L2}(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x} \right)^{1/2} = \frac{1}{\sqrt{A_\Omega}} \|\mathbf{q}_{L2}\|_2. \tag{5.6}$$

Note that since Eq. 5.6 is a constant multiple of the optimal quadratic transport cost, the triangle inequality[1] is satisfied; there is no other transport field satisfying Eq. 5.3 and the boundary conditions that obtains a lower RMSTE than $\mathbf{q}_{L2}$.

The RMSTE, Eq. 5.6, can be seen to penalize the transport itself, while the RMSE, Eq. 5.2, penalizes the bed level changes and thus, according to Eq. 5.3, the divergence of the transport. As a measure of volume times displacement per unit surface area, the RMSTE has units m^2, while the RMSE, which measures a volume per unit surface area, has units m.

Redistributing sediment from h_1 to h_2 through $\mathbf{q}_{L2}$ implies removing sediment at locations for which $\delta = h_1 - h_2 > 0$ and adding sediment at locations for which $\delta = h_1 - h_2 < 0$. We can express the sediment surplus as an excess height[2] given by $\delta_1 = \max(\delta, 0)$ and the sediment shortage as a deficit height[2] given by $\delta_2 = \max(-\delta, 0)$. Of course, in Eq. 5.3, $h_1 - h_2 = \delta_1 - \delta_2$. With these definitions, RMSTE can be seen to measure the smallest overall volume transport of sediment required to excavate δ_1 and fill δ_2. For a zero bias between predictions and observations and boundaries closed for sediment, δ_1 will be transported to δ_2. More in general, sediment may also be added or removed through the boundaries depending on the transportation cost.

5.2.3 Solving the Effective Transport Difference

The solution of Eq. 5.5 proves to be irrotational (see Appendix 5.A) and can therefore be represented as the gradient of a scalar field, the potential ϕ:

$$\mathbf{q}_{L2} = \nabla \phi. \tag{5.7}$$

With Eq. 5.7, Eq. 5.3 can be written as:

$$\nabla^2 \phi = h_1 - h_2. \tag{5.8}$$

Equation 5.8 is a standard Poisson equation for which numerous efficient solvers are available. With ϕ defined through Eq. 5.7, this Poisson equation is fully equivalent to Eq. 5.5. This realization is analoguous to the interpretation of Moser's coupling in terms of optimization theory in Brenier (2003, Section 2.6), where it is

[1] A function $m(A, B)$ is a metric if it is symmetric $m(A, B) = m(B, A)$, positive-definite $m(A, B) \geqslant 0$ and $m(A, B) = 0 \iff A = B$, and satisfies the triangle inequality $m(A, B) + m(B, C) \geqslant m(A, C)$. Both the RMSE and the RMSTE satisfy these criteria.

[2] Surplus or shortage volume (m^3) per squared meter of domain area (m^2), hence an excess or deficit height (m).

shown that the solution to a variant to the L^2 Monge–Kantorovich problem can be represented as a potential flow satisfying the Laplace equation.

The two typical boundary conditions for our application are:

1. Neumann-type boundary condition for a boundary *closed* for sediment:

   $$\mathbf{q} \cdot \mathbf{n} = \nabla\phi \cdot \mathbf{n} = 0 \tag{5.9}$$

   for all points on the boundary $\partial\Omega$. Here $\cdot$ denotes the inner product and the normal vector $\mathbf{n}$ is the unit vector that is perpendicular to the surface $\partial\Omega$ and points outwards from $\partial\Omega$.

2. Dirichlet-type boundary condition for a *free* boundary that allows for sediment transport across the boundary:

   $$\phi = 0 \tag{5.10}$$

   for all points on the boundary $\partial\Omega$, which signifies that there is no constraint on $\mathbf{q}$ on the boundary (see Appendix 5.A).

One can prove that for the case of Dirichlet boundary conditions the solution to Poisson's equation always exists and is unique. The same holds for combination boundary conditions, which consist of Dirichlet boundary conditions on part of the domain boundary and Neumann boundary conditions on the remainder of the domain boundary. For the case of Neumann boundary conditions, a solution only exists if the transport $\nabla\phi \cdot \mathbf{n}$ integrated over the boundary $\partial\Omega$ is consistent with $h_1 - h_2$ integrated over Ω. Then the solution is unique up to an overall additive constant. Clearly, it is sufficient to determine ϕ up to an arbitrary additive constant, which has no impact on the value of the sediment transport $\mathbf{q} = \nabla\phi$.

The above implies that if all boundaries are closed to sediment, the condition for existence of a solution to Eq. 5.8 is that the domain-averaged value of the right-hand-side $h_1 - h_2$ equals zero. The term $h_1 - h_2 = \delta_1 - \delta_2$ was seen to act as a source and sink term, with δ_1 the excess height, which needs to be removed, and δ_2 the deficit height, which needs to be supplied (see Sect. 5.2.2). If its domain-averaged value is equal to zero, hence, in the absence of a bias, the excess sediment δ_1 suffices to fill the sediment deficit δ_2, such that a solution can be found within the domain. If its domain-averaged value is unequal to zero, i.e. in the case of a bias, a net sediment import or export is required to obtain a match between predictions and observations. Then, at least one free boundary condition should be applied.

In specifying the boundary conditions for a particular application, one must bear in mind that $\mathbf{q}$ represents the optimal cumulative transport through which a perfect match is obtained between predictions and observations, hence a transport difference or error. This error is just as arbitrary on the boundary as within the domain, such that in general $\mathbf{q} \cdot \mathbf{n}$ on the boundary $\partial\Omega$ is unknown and there should be no constraint on $\mathbf{q}$ on the boundary. This means that from a physical point of view, *free* boundaries, which are part of the transport optimization, are generally the logical choice. Only in the special case of a boundary that is physically closed for sediment, one may assume that the error $\mathbf{q} \cdot \mathbf{n}$ on the boundary is known and zero. Land boundaries or boundaries beyond the depth of closure (Hallermeier, 1980) may typically be regarded as closed boundaries.

In general, with one or more free boundaries, sediment may be imported or exported through the boundaries depending on the transportation cost. This may be the case both with and without a bias between predictions and observations. In this respect, our method differs from usual optimal transport methods that do not allow a transport across the boundaries and assume that the total mass is contained within the domain.

We refer to the above algorithm for computing an optimal sediment transport as Effective Transport Difference (ETD), since we resolve a transport field that is fully effective in causing morphodynamic change. An arbitrary transport field, satisfying Eq. 5.3, can be decomposed into a rotation-free and divergence-free part (Helmholtz decomposition). Only the rotation-free part, which contains the information about the divergence, results in bed-level changes through Eq. 5.3. Thus, the irrotational, optimal transport field from the least-squares optimization (Eq. 5.5) only contains information that can unambiguously be derived from the bed-level differences and boundary conditions.

5.2.4 Numerical treatment

Section 5.2.3 presented a partial differential equation approach to obtain the quadratic optimal transport $\mathbf{q}_{L2}$. We have implemented this approach using the functions from the Matlab Partial Differential Equation (PDE) Toolbox, which employs a Finite Element Method (FEM) solver for problems on an unstructured grid (Mathworks, 2015). For now, our implementation has been targeted to relatively simple 2D cases, such as shown in this paper (Sects. 5.3.2 and 5.4).

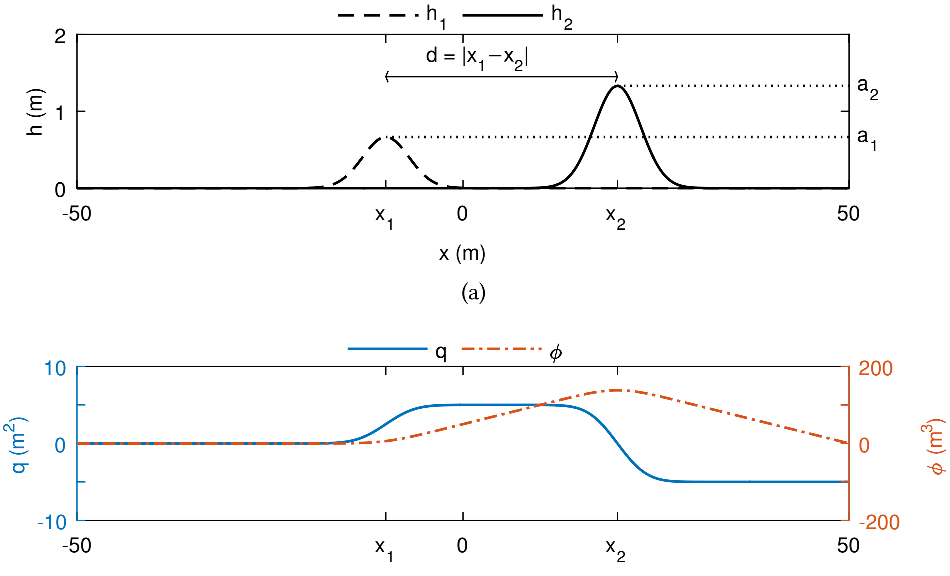
The complex geometry, as required by the PDE toolbox, is generated starting with a rectangular domain from which any “dry points”, representing, for instance, barrier islands, are excluded (see Sect. 5.4). The boundary enclosing the complex geometry, is subdivided into multiple segments, for which a choice between free or closed boundary conditions is available. The Poisson equation is solved on a triangular Delaunay mesh, which is step-wise refined until the solution converges.

5.3 Simple cases

In this section, we compare the behaviour of the RMSTE and RMSE for simple 1D and 2D cases. First, the 1D cases in Sect. 5.3.1 show that the RMSTE does not suffer from the limitations of the RMSE, viz. insensitivity to misplacement distance and the double penalty effect. Next, Sect. 5.3.2 confirms these conclusions based on a simple 2D example. Section 5.3.2 also discusses the characteristics of the 2D optimal transport field and potential for both free and closed boundaries.

5.3.1 Metric behaviour in 1D

The 1D predictions, h_1, and observations, h_2, are represented by equally wide, Gaussian-shaped humps ($\sigma = 3$) on an otherwise flat bed, with amplitudes and centre points a_1 and x_1 and a_2 and x_2, respectively, such that the misplacement distance is $d = |x_1 - x_2|$ (Fig. 5.1a). Our aim is to compare the behaviour of the RMSTE and RMSE for varying misplacement distance d and amplitude ratio $a_1 a_2^{-1}$. For $x_2 = 0, 10, 20$ and 35 m, x_1 was varied such that the predictions were positioned everywhere in the domain. The observed amplitude a_2 was fixed at 1.33 m, while the amplitude of the predictions a_1 varied as $0 \leqslant a_1 a_2^{-1} \leqslant 2$. When the misplacement distance d is larger than the feature width and $a_1 a_2^{-1} = 0$, the classic double penalty case is obtained.



(a)

(b)

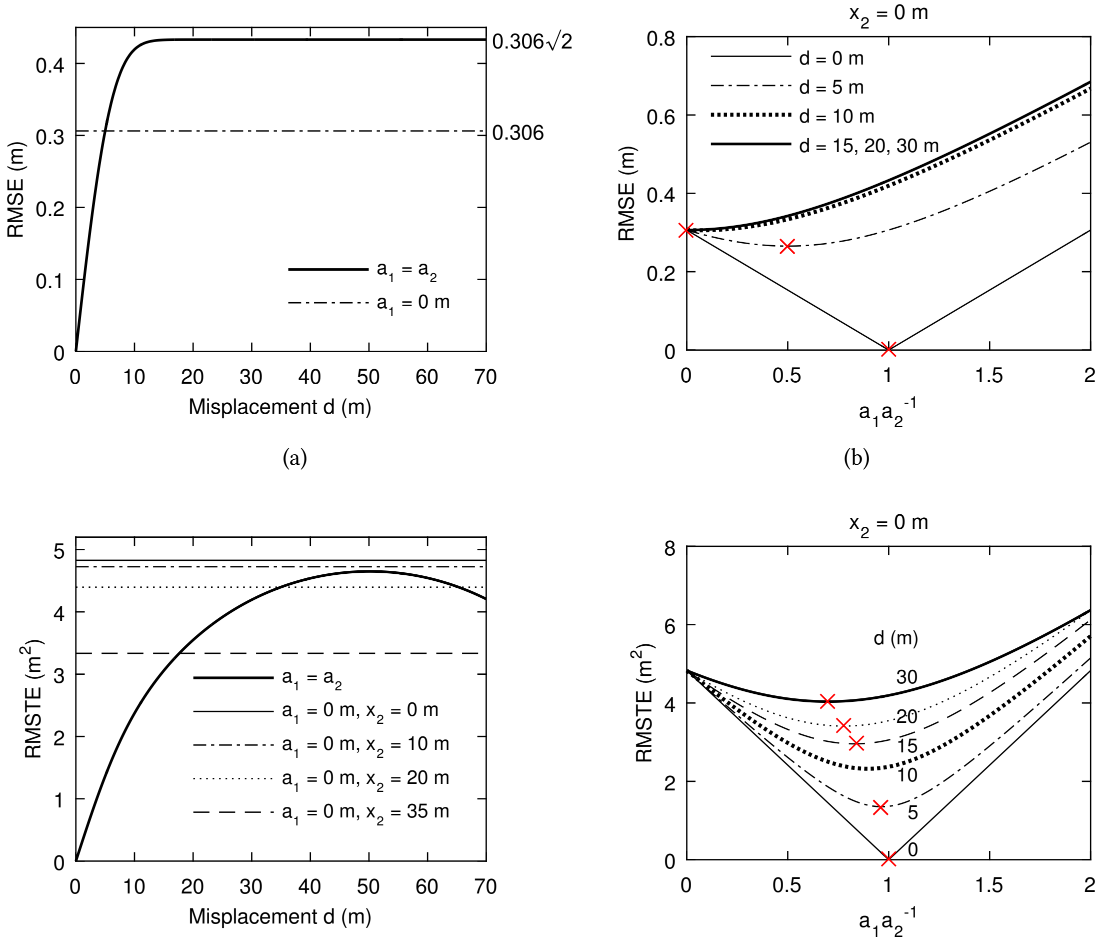
Figure 5.1: Lay-out of the 1D cases with an example solution: **(a)** predicted and observed bathymetries, h_1 and h_2 respectively, consist of equally wide Gaussian shaped humps ($\sigma = 3$) on a flat bed with amplitudes and centre points a_1 and x_1 and a_2 and x_2, respectively and the distance between the two bathymetries $d = |x_1 - x_2|$ (depicted is $a_1 = 0.67$ m, $x_1 = -10$ m, $a_2 = 1.33$ m and $x_2 = 20$ m), **(b)** the corresponding optimal transport $q(x)$ and potential $\phi(x)$ with $\phi(-50) = \phi(50) = 0$ (free boundaries) and, thus, the domain-averaged transport $\overline{q} = 0$.

Obviously, in order to compute the RMSTE, first $q(x)$ needs to be solved. In 1D, if the transport is known at one of the boundaries, for instance $q = 0$, the volume balance has no excess degrees of freedom and only one solution exists, which is found by straightforward numerical integration. This is equivalent to solving the 1D Poisson equation with a closed boundary ($q = 0$) at one end of the domain and an unconstrained boundary ($\phi = 0$) at the other end. Only if the transport is unknown at the boundaries, there is (some) room for optimization in 1D. Considering that the optimal transport q_{L2} is given by the gradient of the potential, the unconstrained condition $\phi = 0$ at either boundary of a 1D domain implies that the transport integrated over the domain is zero, and, thus, the average transport $\overline{q} = 0$. The optimal solution is therefore easily found by integration of the volume balance, while requiring $\overline{q} = 0$. The addition of any nonzero constant to the optimal transport, although still satisfying the volume balance, would increase the transport cost without contributing to bed level changes.

In the presence of a bias, when $\overline{h}_1 \neq \overline{h}_2$, at least one boundary should be free for a solution to exist. For the examples in this section, we have used free boundary conditions at both ends of the domain. Since the free boundary is less constrained than the closed boundary, this will always result in the smallest transport cost.

Figure 5.1b depicts $\phi(x)$ and $q(x)$ corresponding to h_1 and h_2 as shown in Fig. 5.1a. The transports are defined positive in positive x-direction and negative in negative x-direction. The potential $\phi(x)$ is zero at the boundaries and increases with x for positive transports and decreases with x for negative transports, its slope representing the transport magnitude. Obviously, q increases where sediment needs to be eroded and decreases where it needs to be deposited, with the changes in q equal to the volume changes. The average transport $\overline{q} = 0$. The bias requires sediment to be imported, which is, because of the positon of the features relative to the boundaries, most cost-efficiently done from the right boundary only, towards the observed feature, such that $q\,(-50) = 0$. All sediment contained in the excess height $\delta_1 = \max\,(h_1 - h_2, 0)$ of the predicted, left hump is moved to the right for the benefit of the larger deficit height $\delta_2 = \max\,(h_2 - h_1, 0)$ of the observed, right hump.

Figures 5.2a and 5.2c show the RMSE and RMSTE, respectively, as a function of feature misplacement d. Figure 5.2a illustrates that for two equally sized features ($a_1 = a_2$), the RMSE rapidly increases with increasing d, until, when d is larger than the feature width, the RMSE attains a constant value. This value is a factor $\sqrt{2}$ larger than the RMSE for a flat bed, since the double penalty on the MSE translates to the RMSE as a factor $\sqrt{2}$. In contrast, the RMSTE shows an increase with increasing misplacement distances d, until at relatively large d the proximity of the boundaries forces RMSTE to decrease (Fig. 5.2c); at smaller feature spacings, δ_2 is almost fully replenished by δ_1, whereas at larger feature spacing, it becomes more favourable to also export and import sediment in order to excavate δ_1 and



(a)

(b)

(c)

(d)

Figure 5.2: Behaviour of RMSE and RMSTE, for free boundaries, for the 1D cases introduced in Fig. 5.1: **(a)** RMSE as a function of misplacement distance d for predictions with a correct amplitude and for a flat bed prediction, **(b)** RMSE as a function of amplitude ratio $a_1 a_2^{-1}$ for various misplacement distances d and the centre position of the observations at $x_2 = 0\,\mathrm{m}$, **(c)** RMSTE as a function of d for correctly predicted amplitudes as well as for flat bed predictions compared to observations at various centre positions x_2, and **(d)** RMSTE as a function of $a_1 a_2^{-1}$ for various misplacement distances d and the centre position of the observations at $x_2 = 0\,\mathrm{m}$. The red crosses in **(b)** and **(d)** indicate the minima.

fill δ_2, respectively. Note that for closed boundaries RMSTE is strictly increasing with d (not shown).

For equally sized features, $a_1 = a_2$, we have $q(-50) = q(50)$. As a consequence the RMSTE depends on the misplacement distance d only, regardless of the values of x_1 and x_2. The RMSTE for the flat bed prediction strongly depends on the position of the observed hump relative to the boundary and, hence, on x_2, since the entire deficit height δ_2 must be imported. It follows from Fig. 5.2c that for not too large x_2 and d, the RMSTE is larger for the missed feature than for the misplaced feature.

Figure 5.2b confirms that the RMSE rewards an underprediction of the feature amplitude. For a feature, misplaced over a distance smaller than its width, RMSE is minimized for values of $0 < a_1 a_2^{-1} \leqslant 1$. For misplacements larger than the feature size, the flat bed prediction, $a_1 = 0$, receives the smallest RMSE. Although the RMSTE also has minima at values of $a_1 < a_2$ for $d > 0$, these minima appear at values of $a_1 a_2^{-1}$ relatively close to 1 (Fig. 5.2d). Note that Figs. 5.2b and 5.2d are valid for $x_2 = 0$.

The above demonstrates that: (1) the RMSTE, as opposed to the RMSE, is able to account for misplacement distance; and (2) that the double penalty effect is specific to the RMSE. Whether or not the RMSTE is larger for a flat bed prediction than for a correctly sized but misplaced feature depends strongly on the situation.

5.3.2 Demonstration for simple 2D case

In this section, we present a simple example to illustrate the behaviour of the RMSTE in 2D and to provide insight in the characteristics of the 2D potential and optimal transport fields, for various boundary conditions.

Figure 5.3 compares an observed 2D feature with three suboptimal predictions: (1) a flat bed prediction, (2) a misplaced feature, and (3) a misplaced feature at a larger misplacement distance. The (R)MSE and RMSTE error values are given in Table 5.1. The RMSTE is computed with three different sets of boundary conditions: free boundaries only, closed boundaries only (not applicable for a bias) and a combination of a closed South boundary and free boundaries elsewhere (further on referred to as combination boundaries). The different boundary conditions result in the same ranking of the three predictions. However, fewer constraints lead to lower transport costs, such that for all predictions the lowest RMSTE is obtained for free boundaries.

Prediction	MSE	RMSE	RMSTE$_{\text{free}}$	RMSTE$_{\text{combination}}$	RMSTE$_{\text{closed}}$
	($\times 10^{-2}$m^2)	($\times 10^{-1}$m)	($\times 10^{-2}$m^2)	($\times 10^{-2}$m^2)	($\times 10^{-2}$m^2)
1	1.05	1.03	1.04	1.08	n.a.
2	2.11	1.45	0.87	0.90	0.94
3	2.11	1.45	1.11	1.23	1.32

Table 5.1: (R)MSE and RMSTE with free, combination (only South boundary closed) and closed boundaries, for predictions 1, 2 and 3.

Of course, predictions 2 and 3 are diagnosed with an MSE and RMSE that are larger by a factor 2 and $\sqrt{2}$, respectively, than for prediction 1; prediction 2 and 3 are penalized twice, both where the predicted feature is and where it should be,

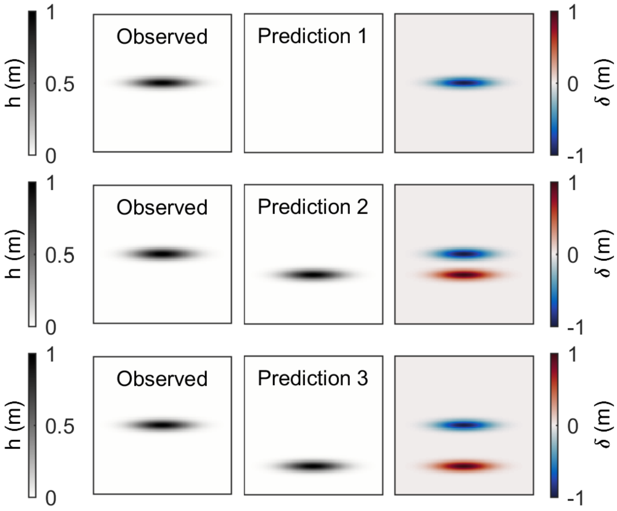


Figure 5.3: Three alternative predictions of the same observed feature. *Top panels*: the featureless prediction 1 has a nonzero difference δ between predicted and observed depth values at the location of the observed feature only. *Middle panels*: prediction 2, which reproduces the feature at the wrong location, is penalized twice, since δ is nonzero both where the predicted feature is and where it should be. *Lower panels*: prediction 3, with a larger misplacement distance, is also penalized twice.

whereas prediction 1 is penalized at the location of the observed feature only (see Fig. 5.3). As opposed to the RMSE, the RMSTE distinguishes between prediction 2 and 3, the feature with the smaller misplacement distance (prediction 2) receiving the lower RMSTE. Prediction 2 also outperforms the flat bed prediction (prediction 1), while the feature with the larger misplacement distance (prediction 3) obtains the worst score.

Evidently, in line with the findings for the 1D cases (Sect. 5.3.1), for correctly sized features, the RMSTE increases with misplacement distance, until, in the extreme, sediment exchanged across the model boundaries may lead to a lower RMSTE. As discussed in Sect. 5.3.1, whether or not a misplaced feature outperforms a missed feature is determined by the (optimal transport cost for) the considered morphological patterns and, hence, depends on the boundary conditions, the size and shape of the observed and misplaced features and their position relative to each other and to the domain boundaries.

Figure 5.4 illustrates the characteristics of the potential and optimal transport for prediction 3, using closed boundaries. The left panel shows the optimal transport field moving sediment from the excess height δ_1 to the deficit height δ_2, hence from the red to the blue patches, at minimum cost. The transport field $\mathbf{q} = \nabla\phi$ is fully determined by the potential ϕ (bottom right panel), given by Poisson's Eq. 5.8. Thus, the transport occurs everywhere at right angles to the equipotential lines, i.e. the lines of constant ϕ, and the spacing of the equipotential lines reflects the transport magnitude (top right panel), from which RMSTE is easily computed. At the closed boundaries, the equipotential lines are perpendicular to the boundaries, corresponding to zero transport through the boundaries. Naturally, the quadratic cost function governs the transport pattern. The transport magnitudes at different locations are weighted quadratically, so extremes are heavily penalized. This leads to the observed somewhat diffuse transport pattern with curved transport pathways.

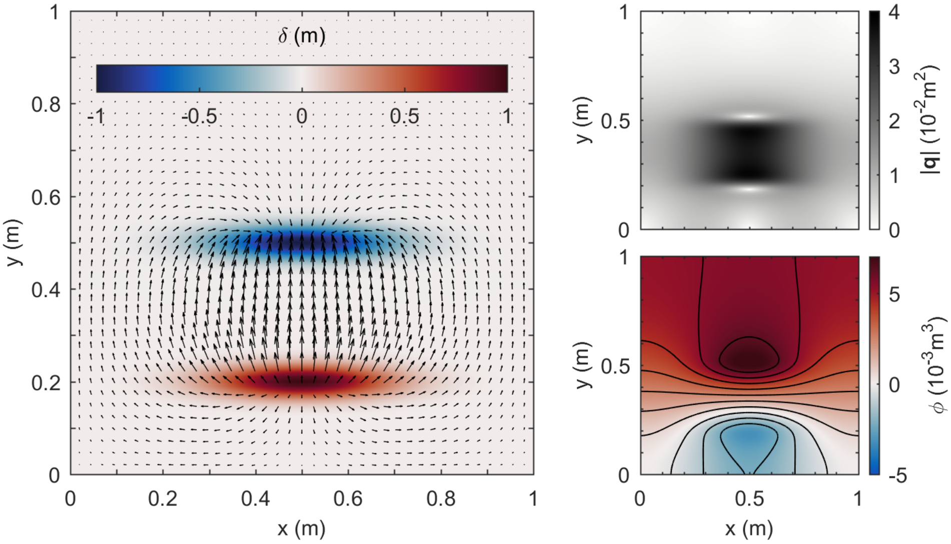


Figure 5.4: Optimal solution, with closed boundaries, for prediction 3. *Left panel*: bed level difference $\delta = h_1 - h_2$ with the arrows indicating the transport field (length and direction of arrows indicative of the transport magnitude and direction, respectively). *Right panels*: transport field represented by the transport magnitude (*top*) and potential ϕ (*bottom*).

Figure 5.5 shows the transport magnitude and potential for the situation of free boundaries all around as well as for combination boundaries, which combine a closed South boundary with free boundaries elsewhere. One can verify that the potential is zero at free boundaries allowing transport across the boundaries. The transport magnitudes are smallest when all boundaries are free, resulting in the lowest RMSTE (see Table 5.1). Unless there is additional knowledge about the error on the boundaries, for instance for a boundary physically closed to sediment, the use of free boundaries is advised (see Sect. 5.2.3).

5.4 Example of a tidal inlet

In this section, we test and illustrate the RMSTE for a more realistic case of a tidal inlet. We diagnose the correspondence between multiple pairs of morphological fields, generated by Delft3D, as well as the relative ranking between the

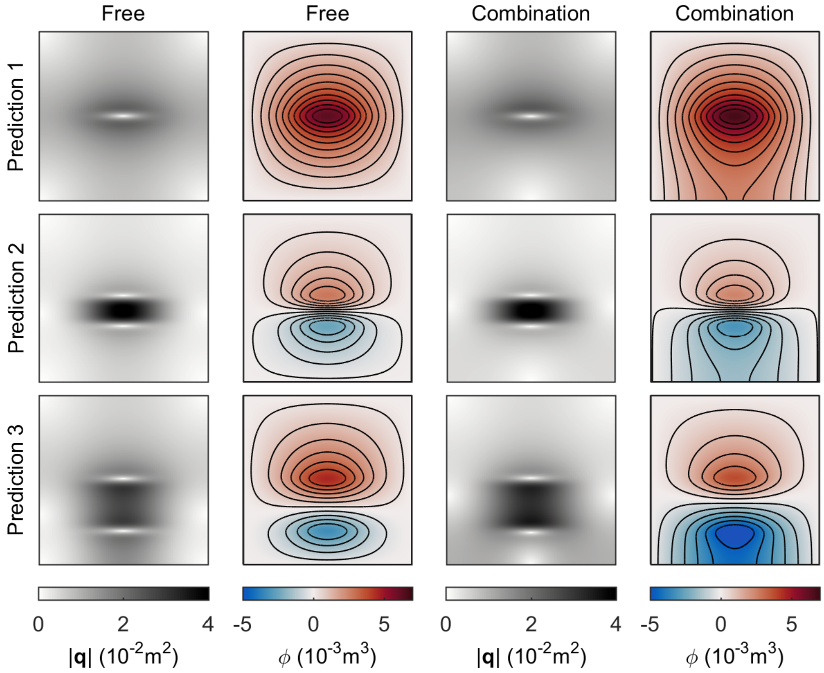


Figure 5.5: Transport magnitudes and potential for predictions 1, 2 and 3 for free boundary conditions and combination (South boundary closed, remainder free) boundary conditions.

pairs. First, an overview of the model runs and morphological fields is given in Sect. 5.4.1. Next, in Sect. 5.4.2, we test the behaviour of the RMSTE for fields with misplaced tidal channels due to incorrect Coriolis settings. Subsequently, Sect. 5.4.3 presents a full comparison demonstrating the differential behaviour of the RMSE and RMSTE.

5.4.1 Overview

Starting from an initially highly schematized tidal inlet (Fig. 5.6), we have generated ten morphological fields with Delft3D. The inlet geometry and boundary forcing are chosen such as to resemble the Wadden Sea inlet of Ameland. The tidal basin is rectangular with an area of 15 × 10 km^2 and a uniform initial depth of 2 m; the entrance has a width of 2 km, and the seabed initially slopes from −2 m at the barrier islands to −10 m at the offshore (Northern) boundary (Roelvink, 2006). The model has a uniform grid size of 100 × 100 m^2.

For the base run O (see Table 5.2), the latitude was set to 0° and a uniform, harmonic water level variation was applied along the offshore boundary with a period of 12 h and a water level amplitude a of 1 m. The standard sediment transport formulations according to van Rijn were applied, with a multiplication factor for the suspended sediment reference concentration f_{sus} = 1 and a median sediment size

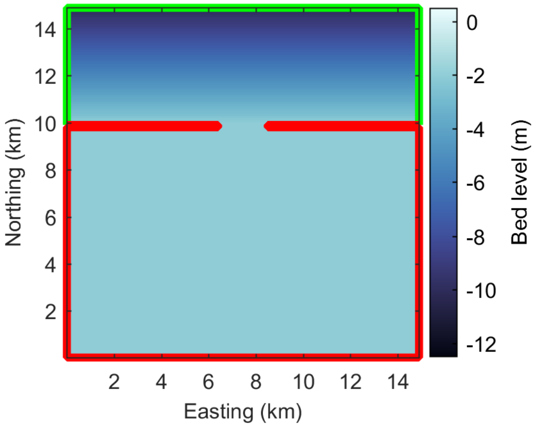


Figure 5.6: Initial bathymetry with the free sea boundaries for Poisson's Eq. 5.8 in green and the closed land boundaries in red.

D_{50} = 200 μm. The other 9 runs listed in Table 5.2 are variations to the base run O with respect to latitude, f_{sus}, D_{50}, tidal amplitude a and tidal direction. The latter was changed, for run A only, from cross-shore to alongshore by applying a phase difference along the Northern boundary. The final bathymetries of the 10 runs are shown in Fig. 5.7. For the computation of the ETD and, subsequently, the RMSTE between pairs of depth fields (Sects. 5.4.2 and 5.4.3), we have considered the land boundaries and sea boundaries as closed and free boundaries, respectively (see Fig. 5.6).

Run	Latitude	Amplitude (m)	Direction	f_{sus}	D_{50} (μm)
O	0°	1.0	C	1.0	200
A	0°	1.0	L	1.0	200
B	0°	0.67	C	1.0	200
C	0°	1.5	C	1.0	200
D	0°	0.5	C	1.0	200
F	90° N	1.0	C	1.0	200
G	90° S	1.0	C	1.0	200
L	53° N	1.0	C	1.0	200
M	53° N	1.0	C	1.5	200
N	0°	1.0	C	1.0	250

Table 5.2: Overview of the 10 runs used to generate the morphological fields of Fig. 5.7. O is the base run, the others are variations with respect to latitude, tidal amplitude and direction [C(ross)- or L(ongshore)], transport parameter f_{sus} and D_{50}. The labels are chosen such as to be consistent with Bosboom and Reniers (2014b, i.e. Ch. 4) and Mol et al. (2015).

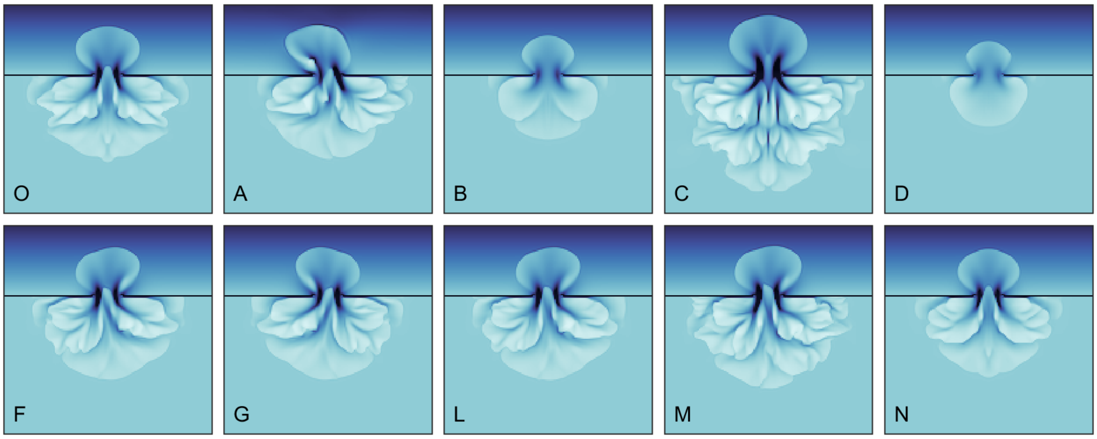


Figure 5.7: Final bathymetries of the 10 runs with settings according to Table 5.2. The horizontal and vertical axes and the color scaling are as in Fig. 5.6.

5.4.2 Variation in Coriolis

First, in this section, the model-generated depth fields O, F, G and L (see Table 5.2 and Fig. 5.7) are considered, which only differ with respect to the latitude, and, hence, Coriolis parameter. We label depth field L, the one with 53° N, as the observations and regard the other three as three competing predictions. The pairs of computations and observations are named by the label of the predictions followed by the label of the "observations" (see Table 5.3).

Figures 5.8 and 5.9 show the bed level differences $\delta = h_1 - h_2$ and transport magnitudes $|\mathbf{q}|$, respectively. The required transport corrections are mostly confined to the flood and ebb tidal delta areas, with zero or small transports outside these delta regions. Sediment is relocated from the excess locations to the shortage locations, i.e. from the red to the blue patches in Fig. 5.8. On the ebb tidal delta, this results in a transport between the delta flat and the outer edges in both directions. In the flood tidal delta, sediment is transported to locations where channels are wrongly predicted and away from locations where they should have been predicted. The transport distances are limited, since the Coriolis errors require only local corrections to feature locations.

Both the RMSE and RMSTE increase with increasing latitude deviation (Table 5.3), since both the misplaced volumes of sediment and the misplacement distances increase with latitude error. Under these circumstances, the various error metrics, including the metrics based on the field deformation or image warping method (see Bosboom and Reniers, 2014b), demonstrate the same qualitative behaviour.

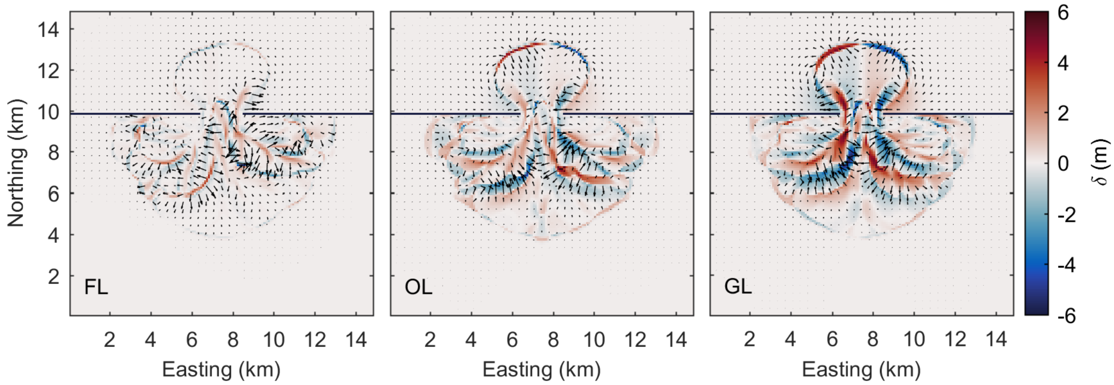


Figure 5.8: Bed level differences and transport fields, with the length and direction of the arrows indicative of the transport magnitude and direction, respectively, for cases FL, OL and GL (see Table 5.3).

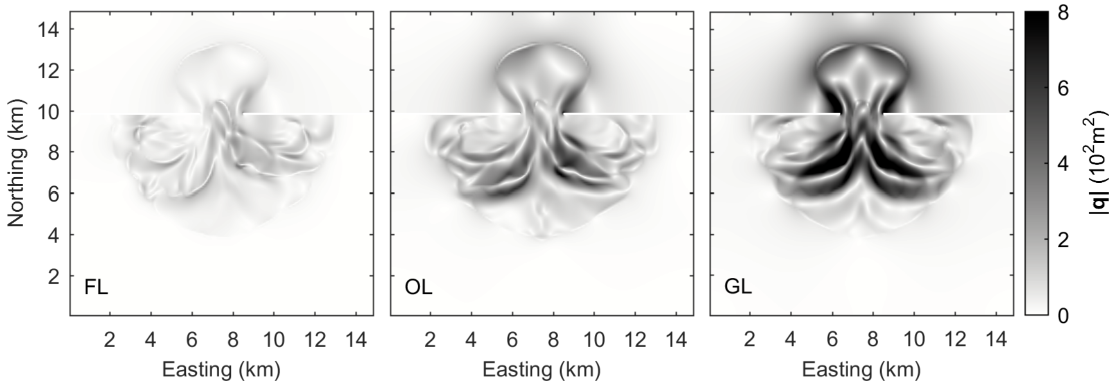


Figure 5.9: Transport magnitudes for cases FL, OL and GL with closed land boundaries and free sea boundaries.

Case	Latitude model	Latitude observed	RMSE (m)	RMSTE (×10^2m^2)
FL	90° N	53° N	0.29	0.5
OL	0°	53° N	0.52	1.2
GL	90° S	53° N	0.73	2.0

Table 5.3: RMSE and RMSTE for three cases with errors in latitude and hence Coriolis parameter. Case names consist of the label of the predictions followed by the label of the "observations", which are taken as the model outcome at 53° N.

5.4.3 Comparison of all fields

In this section, the predictions A to N are compared to the observations O by means of the RMSE and the RMSTE (see Table 5.4). From Table 5.4, it is clear that the RMSE and RMSTE lead to a different ranking amongst the predictions, with prediction L receiving the lowest RMSTE and prediction N the lowest RMSE. Further, as opposed to the RMSE, the RMSTE is seen to discriminate between predictions F (or its mirrored prediction G) and M as well as between A and B. The distinctive behaviour of the two error metrics is a logical consequence of their different definition. Below, we highlight and explain some of these differences on the basis of the underlying fields of bed level differences and transports (Figs. 5.10 and 5.11).

Case	RMSE (m)	RMSTE ($\times 10^2 m^2$)
AO	0.78	2.6
BO	0.77	5.5
CO	1.16	8.2
DO	0.95[3]	7.8
FO	0.59	1.4
GO	0.59	1.4
LO	0.52	1.2
MO	0.59	2.2
NO	0.47	1.8

Table 5.4: RMSE and RMSTE for predictions A to N compared to "observations" O, the model outcome at 0°, hence without the influence of Coriolis.

Prediction L, which has a modelled latitude of 53° N rather than the "real" 0°, is awarded the lowest RMSTE. Predictions F and G, with a 90° modelled latitude, receive the second best RMSTE. The wrongly predicted Coriolis deflection leads to a distortion of the outer edges of the ebb tidal delta and a mispositioning of the channels on the flood tidal delta and in the inlet gorge. The sediment transport required to correct these Coriolis errors takes place over short distances only, explaining that, measured by the RMSTE, predictions F, G and L outperform the other predictions, including prediction N, which is the best prediction in terms of RMSE. The too high grain size of the latter prediction results in an underdeveloped delta, which must be corrected by transporting sediment from the channel locations to build the flats and extend the delta rims. The relatively large distances over which this sediment is transported explains the larger RMSTE compared to the predictions with Coriolis error, even though, based on the RMSE, the amount of misplaced sediment is smaller. This underlines again that the RMSE measures misplaced sediment volumes only, whereas the RMSTE takes misplacement distance into account as well.

Despite receiving the same values of RMSE, predictions F and M behave differently in terms of RMSTE. The erroneous Coriolis deflection that both predictions suffer from, is stronger for prediction F than for prediction M. Prediction M, how-

[3] Corrected from 0.96 as previously listed in Bosboom and Reniers (2014b).

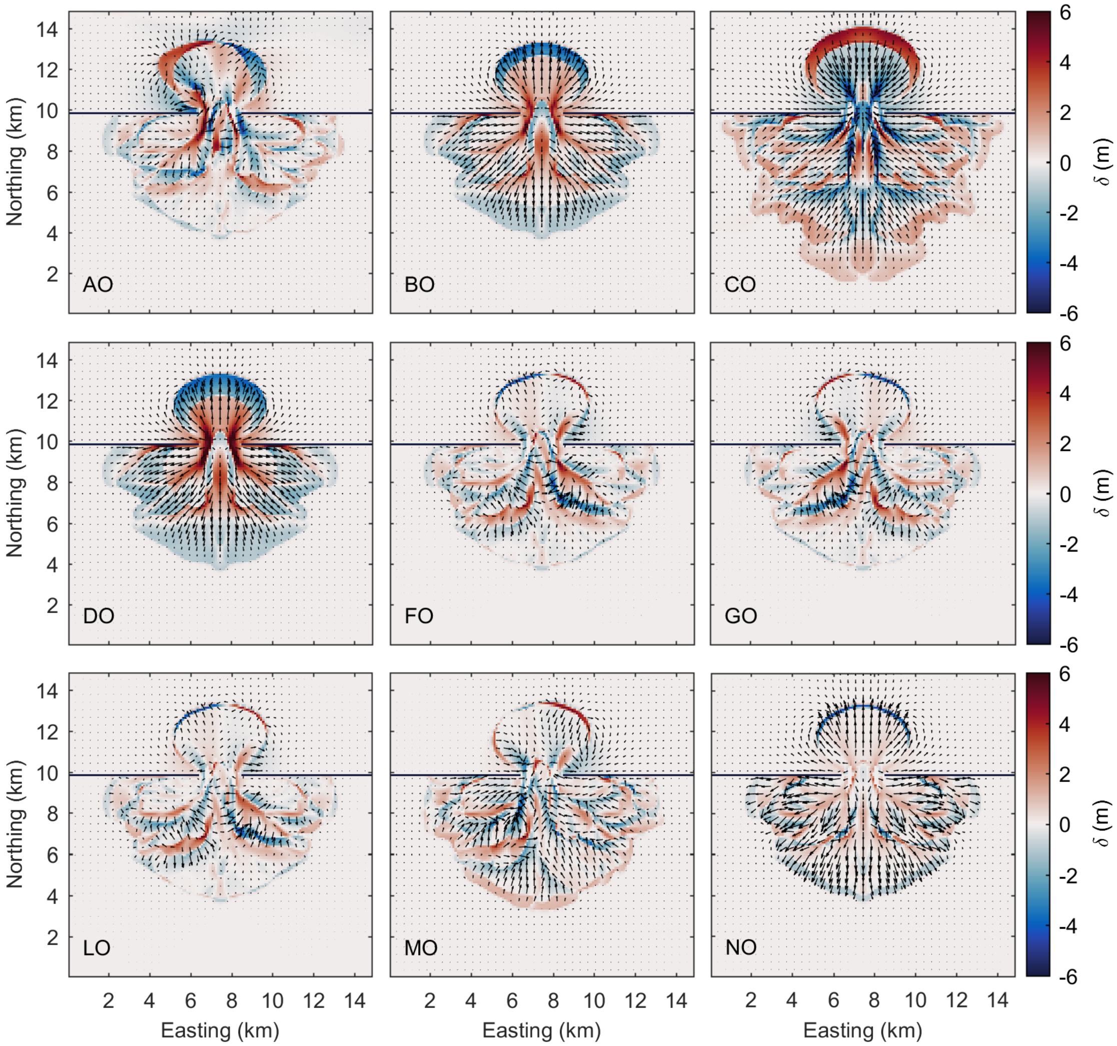


Figure 5.10: Bed level differences and transport fields, with the length and direction of the arrows indicative of the transport magnitude and direction, respectively, for predictions A through N compared to the "observations" O.

ever, has an additional error source that requires a corrective transport over larger distances; due to too large suspended sediment transports (f_{sus} = 1.5 instead of 1), the inlet system is overdeveloped. The corrective transport pattern for prediction M shows the two error sources operating at different spatial scales, of which the longer scales weight heavier towards the RMSTE. The result is a domain-averaged corrective sediment transport that is larger for prediction M than for F.

Predictions A to D were added to allow a comparison with the $\mathrm{RMSE}_{\mathrm{w}}$, the combined error metric based on the field deformation or image warping method of Bosboom and Reniers (2014b). The $\mathrm{RMSE}_{\mathrm{w}}$ combines all relevant information

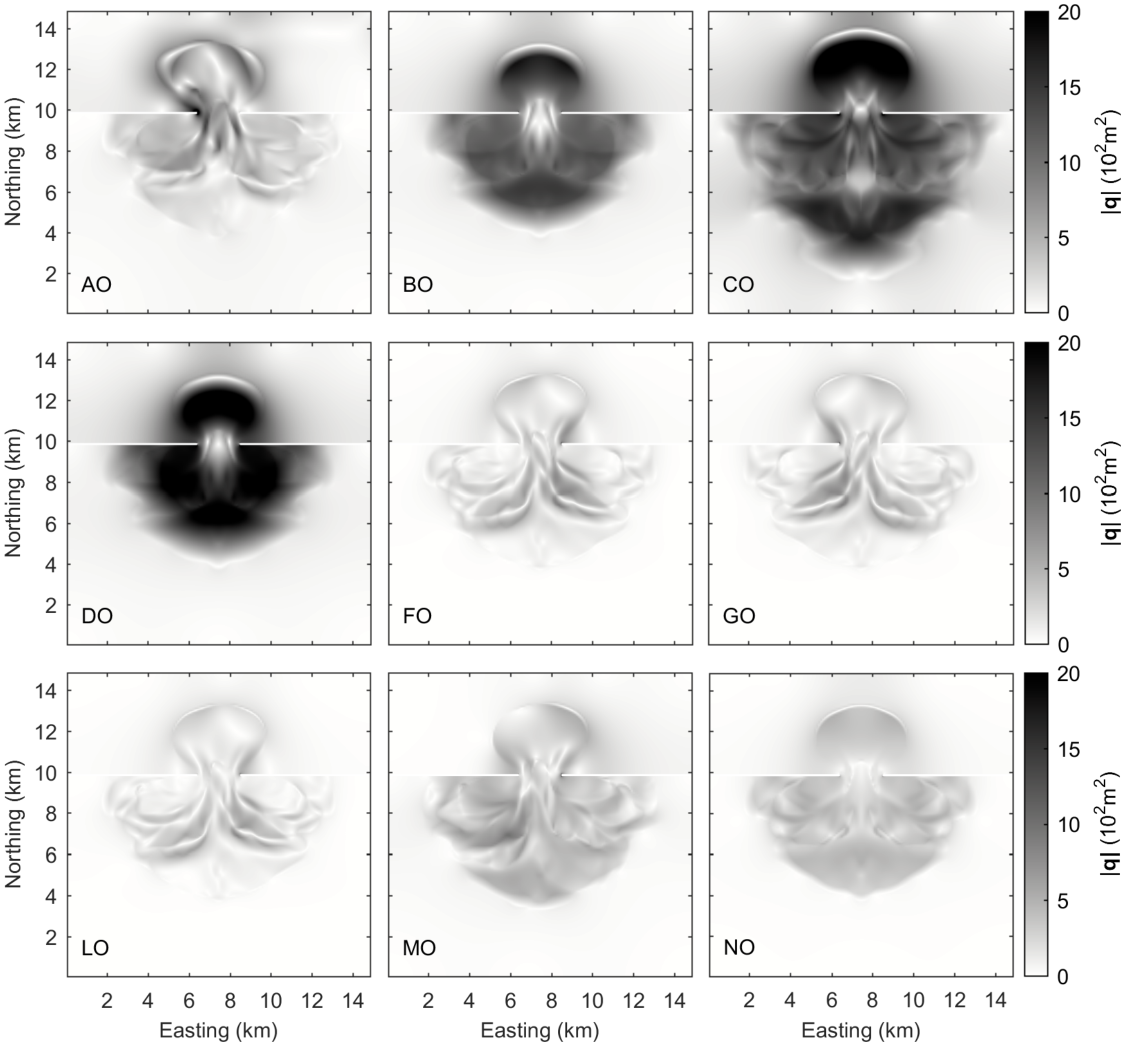


Figure 5.11: Transport magnitudes for predictions A through N compared to the "observations" O.

on location errors and pre- and post-warp intensity (i.e. bed level) errors. It depends on a user-defined parameter $D_{\max}$, which represents the maximum distance over which morphological features may be displaced for the prediction to still get (some) credit for predicting these features. Both the RMSTE and the RMSE$_\mathrm{w}$ diagnose prediction A to be a better prediction than B, in spite of the similar values for RMSE (Table 5.5). Whether prediction C or D is diagnosed the better prediction by the RMSE$_\mathrm{w}$ depends on the chosen value for $D_{\max}$. In contrast, the RMSTE does not allow such a parameter. Based on the RMSTE, and hence on the required amount of corrective sediment transport, prediction D outperforms prediction C.

Case	$\mathrm{RMSE_w}$ (m)	$\mathrm{RMSE_w}$ (m)	RMSE	RMSTE
	D_{max} = 3000 m	D_{max} = 1000 m	(m)	($\times 10^2 \mathrm{m}^2$)
AO	0.49	0.63	0.78	2.6
BO	0.60	0.71	0.77	5.5
CO	0.78	1.02	1.16	8.2
DO	0.84	0.94	0.95	7.8

Table 5.5: The combined error metric $\mathrm{RMSE_w}$ from the image warp (with D_{max} = 3000 m and 1000 m) for predictions A to D compared to "observations" O (values from Bosboom and Reniers, 2014b). The values for RMSE and RMSTE are copied from Table 5.4 for ease of reference.

A final remark concerns the free boundary conditions. From Figs. 5.10 and 5.11, it can be seen that for predictions A to D, with larger morphological change closer to the North boundary, there is a small corrective sediment transport across this boundary. Note that these predictions require a net sediment exchange with the outside world due to the presence of a (small) bias. The contribution of the transport across the North boundary to the RMSTE is limited, as can be verified from Fig. 5.11.

5.5 Discussion

Sections 5.3 and 5.4 have shown that the newly introduced RMSTE is capable of discriminating among model results, which is an important requirement of any error metric. We have seen that the RMSTE may lead to a different judgement as to which of two predictions is better than the RMSE, since it highlights other aspects of model performance. The RMSE measures the amount of misplaced sediment, and, hence, penalizes small misplacements of features heavily. As a consequence, it is difficult to demonstrate the quality of a high-variability prediction with the RMSE. The RMSTE on the contrary, is based on the corrective sediment transport from the predicted to the observed morphological field and, consequently, not only takes the amount of misplaced sediment into account, but also the distance over which this sediment is misplaced. Hence, larger spatial scales in the bathymetric error fields, requiring larger corrective transport distances, are penalized heavier than shorter scales. For the simple cases in Sect. 5.3, this was reflected in the RMSTE increasing with the misplacement distance of the considered features and being free from the consistent favouring of flat bed predictions that the RMSE suffers from. Similarly, Sect. 5.4 demonstrated, for more realistic bathymetric patterns, that more localized sediment misplacements, due to, for instance, incorrect Coriolis deflections, are diagnosed with better RMSTE scores than misplacements similar in volume but over larger distances. Section 5.4 further indicates that inspection of the corrective transport fields, underlying the RMSTE, may provide some guidance as to how the model should be improved. As an example, for case

MO in Sect. 5.4.3, the transport pattern revealed two error sources operating on different spatial scales, which, when isolated, may be separately addressed for the improvement of the model.

Based on the above described results, we expect that the RMSTE will enable a more balanced comparison between morphodynamic model predictions. The current validation practice of only using a point-wise accuracy metric—for example the (R)MSE—or a skill score based on such a point-wise metric—for example the mean-squared-error skill score known as Brier skill score (BSS)—tends to reward predictions that underestimate the variability of morphodynamic change (see Bosboom et al. (2014), i.e. Ch. 2, and Bosboom and Reniers (2018), i.e. Ch. 3). This undesirable effect can be counteracted by also taking the RMSTE into account. We further anticipate that the RMSTE will be helpful in calibrating morphodynamic models with respect to the morphodynamic timescale. In a first calibration step, an automated calibration routine, which minimizes the RMSTE, may be able to determine the optimal global model settings, such as certain transport parameters, that merely affect the morphodynamic timescale. In a next step, a more detailed calibration of other parameters can be undertaken using multiple error metrics, amongst others the RMSE and the RMSTE.

In Sect. 5.3.1, we have seen that the proximity of the boundaries may restrict the increase of the RMSTE with the misplacement distance between two features, or even cause the RMSTE to decrease. This can occur when it is cheaper to (partly) export the excess height δ_1 through the one boundary and (partly) import the deficit height δ_2 through another boundary than to directly move δ_1 towards δ_2. It may seem counterintuitive at first, that free boundaries could prevent a predicted and observed feature on either side of the model domain to be (fully) associated with each other. However, the transport error on the boundary is equally unknown as within the domain, such that free boundaries, which are themselves part of the optimization, are generally the logical choice. Only in the special case of a boundary that is physically closed for sediment, such as land boundaries, one may assume that the transport error on the boundary is known and zero. The transport across free boundaries will be relatively small when the bias between predictions and observations is small and the model boundaries are chosen far away from the regions of morphodynamic change, as would generally be the case in practical applications (see also Sect. 5.4). In these cases, also the effect of the boundary conditions on the RMSTE will be small.

As opposed to the RMSE, the RMSTE requires fields to operate on, which complicates its application in data-poor environments. A solution could be to interpolate the data to the computational grid using straightforward interpolation methods. Alternatively, more advanced stochastic models may be used to generate realistic realisations of the seabed, consistent with the available data (Novaczek et al., 2019; Williams et al., 2017). The sensitivity of the RMSTE to data coverage and resolution can be assessed in practice by evaluating the difference between the RMSTE values computed using different methods to estimate missing data. Note that both the RMSE and the RMSTE can be expected to be sensitive to the spatial resolution of the data. In fact, the sensitivity of the RMSTE to spatial resolution is likely to be smaller than of the RMSE, since the first gives more weight to larger spatial scales in the bathymetric error fields than to shorter scales. This also raises the question of the validity of computing the RMSE based on the measurement locations only, which could also be addressed by a sensitivity analysis using multiple realisations of the seabed consistent with the measurements.

In principle, a skill score could readily be derived from the RMSTE, in the same manner as skill scores have been derived from the (R)MSE (Gallagher et al., 1998; Sutherland et al., 2004). Like any skill score, it would inherit the characteristics of the error metric it is based on, in this case the RMSTE, and be critically dependent on the choice of the reference prediction. We expect however that the common choice of the initial morphology as the reference prediction will not be able to create the required level playing field, as was previously demonstrated for the MSESS/BSS (Bosboom et al., 2014; Bosboom and Reniers, 2018).

The ETD expresses the mismatch between predictions and observations in terms of a sediment transport field that is able to transform the predictions to perfectly match the observations. This method, by definition, allows for the redistribution of the excess sediment volume though splitting or coalescing and implies that bed features are not necessarily kept intact. While bed forms are created or flattened out, sediment is redistributed over the morphological scales. The transport direction of sediment contained in a misplaced feature is not necessarily the same as the direction in which the predicted feature needs to be moved. In the examples of Sect. 5.3, the direction of feature movement, for instance from left to right, coincided with the transport direction of the sediment contained in the predicted feature. If we would multiply the bed levels by −1, such that the features were channels rather than humps, the required feature displacement would still be from left to right, but the sediment, rather, would be moved towards the predicted feature, from right to left; the excess height δ_1 and deficit height δ_2 are now found at the location of the observations and predictions, respectively, instead of the other way around. This highlights one of the important differences between the ETD, which moves sediment, and the image warp, such as employed in Ch. 4 (Bosboom and Reniers, 2014b), which, roughly speaking, moves features. This warping method finds an optimal displacement field by minimizing a regular L^2 distance (Eq. 5.1, with $p = 2$). It essentially shifts pixels by (locally) stretching or compressing the morphological pattern to better match the observations. Bed level differences between corresponding features in the predictions and observations prevent an exact match. The combined error metric $\mathrm{RMSE_w}$ as presented in Ch. 4 (Bosboom and Reniers, 2014b) weights both the remaining RMSE after the optimal transformation and the magnitude of the displacements required to obtain this reduced error, according to a user-defined parameter. As a consequence, the RMSE_w can be expected to capture the visual disagreement between morphological patterns, whereas the RMSTE represents the minimum cost, in terms of (squared) sediment transport and, hence, work, to bridge the deviations between the morphological patterns. Advantages of the ETD method over the image warp are that the ETD is mass-conserving, parameter-free and symmetric, the optimal transport from observations to predictions being the inverse of the optimal transport from predictions to observations.

The transport fields $\mathbf{q}_{L2}$, as found in Sects. 5.3 and 5.4, are cumulative, corrective and net transport fields. Here, *cumulative* refers to the time-integration of the transport and *corrective* signifies that the transport fields represent the transport differences between predictions and observations, rather than observed or modelled transports between consecutive moments in time. Further, *net* expresses that they present the "cheapest" way, based on the 2-norm of the transport field, to move the mispredicted sediment volumes to the right locations. Thus, from the multitude of corrective transport fields satifying the volume balance Eq. 5.3, the transport that minimizes the amount of squared work is thought to best represent the mismatch between predictions and observations. Since the optimal transport $\mathbf{q}_{L2}$ is irrotational, it is fully effective in causing morphodynamic change and only contains information that can unambiguously be derived from the bed-level differences and boundary conditions, see Sect. 5.2.3. The physical justification of $\mathbf{q}_{L2}$ as the optimal transport is found in the choice of the cost function formulated in terms of work, rather than in a connection to the usual transport descriptions based on hydrodynamic drivers, which may lead to transport fields that are not optimized with respect to the cost function. Obviously, the exponent p in the cost function can be expected to influence the transport pattern. With our pragmatic choice of $p = 2$, the transport magnitudes at different locations are weighted quadratically, so extremes are heavily penalized. This leads to somewhat smeared out transport patterns with curved transport pathways, as found in Sects. 5.3 and 5.4. With $p = 1$ on the other hand, the transport magnitudes at different locations are weighted proportionally, so the cost function is likely to be less affected by local large transports, which may lead to more pronounced transport patterns.

5.6 Conclusions and perspectives

In this paper, we have presented a novel diagnostic tool for morphodynamic model validation. The employed ETD method solves an optimal transport problem that moves sediment from the one bathmetry (the predictions) to the other bathymetry (the observations) at minimum quadratic transport cost and, thus, work. The quadratic cost function allows a reformulation of the problem in terms of a Poisson partial differential equation, which is uniqely solvable, at least up to an additive constant. A new error metric, the RMSTE, is defined as a constant multiple of the optimal quadratic cost. As such, it measures the error in terms of the net corrective sediment transport volume required for a match with the observations. By penalizing the total sediment transport, the spatial structure of the error is taken into account; the RMSTE is sensitive to the volumes of misplaced sediment as well as to the distance over which this sediment must be transported. Advantages of the ETD method over the image warp of Ch. 4 (Bosboom and Reniers, 2014b) are that the ETD is mass-conserving, parameter-free and symmetric, the optimal transport from observations to predictions being the inverse of the optimal transport from predictions to observations.

The results have shown that the RMSTE, as opposed to the RMSE, is able to discriminate between predictions that differ in the misplacement distance of predicted morphological features. Also, the RMSTE avoids the consistent favouring of the underprediction of the variability of morphodynamic change that point-wise accuracy metrics, such as the RMSE, and the mean-squared-error skill score known as BSS are prone to.

By definition, each error metric condenses a large amount of data into a single number, therewith highlighting certain aspects of morphological model performance only. Therefore, we recommend that a combination of metrics is used in the validation of morphological models and that the weighting is determined by the goal of the simulation. We expect that the addition of the RMSTE enables a fairer comparison between morphodynamic model predictions, by avoiding some of the pitfalls of point-wise metrics and by defining the error in terms of a quantity that is at the heart of morphodynamic model validation.

In future studies, the behaviour of the RMSTE in a range of practical applications will need to be considered. In order to do so, a more robust implementation of the ETD is required in order to deal with arbitrary model domains. Further, we anticipate that valuable additional information can be extracted from the optimal transport fields by isolating the various scales in the transport fields, for instance using our scale-selective validation method of Ch. 6 (Bosboom and Reniers, 2014a).

The choice of $p = 2$ in the optimization problem, leading to quadratic transport costs, has enabled a relatively straightforward solution procedure resulting in a rotation-free optimal transport. For $p = 1$ and a domain boundary closed to sediment, our formulation and the L^p Monge–Kantorovich problem are equivalent and correspond to the original Monge mass transfer, which guarantees the shortest possible weighted transport distance and smallest transport magnitude. Numerical methods for solving the L^1 problem exist (Benamou and Carlier, 2015), but are considerably more complex than our L^2 solution procedure. Nonetheless, it may be worthwile to explore possibilities to solve the L^1 optimization problem. Such an approach would lead to the introduction of a new error metric, the mean absolute transport error (MATE), which can be expected to behave differently than the RMSTE. The MATE is to the RMSTE as the MAE is to RMSE, with that difference that MATE is based on $\mathbf{q}_{L1}$ rather than $\mathbf{q}_{L2}$.

5.A Proof of irrotationality of transport field

Here we prove the claim in Sect. 5.2.3 that the minimizer of

$$\left| \int_{\mathbf{x}\in\Omega} |\mathbf{q}(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x} \right|^{1/2} \tag{5.11}$$

under the constraint of the volume balance (Eq. 5.3) is irrotational. The corresponding Lagrangean $\mathscr{L}$ is

$$\mathscr{L}(\mathbf{q}, \lambda) := \int_{\mathbf{x}\in\Omega} \frac{1}{2} |\mathbf{q}(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x} + \int_{\mathbf{x}\in\Omega} \lambda(\mathbf{x}) \left(\nabla \cdot \mathbf{q}(\mathbf{x}) + h_2(\mathbf{x}) - h_1(\mathbf{x}) \right) \tag{5.12}$$

with λ the Lagrange multiplier for the constraint and $\nabla\cdot$ the divergence operator. Note that the first term is equivalent to Eq. 5.11 as the cost function. At the minimum, the variation of the Lagrangian ($\delta\mathscr{L}$) with respect to $\mathbf{q}$ is zero, hence:

$$0 = \delta\mathscr{L}(\mathbf{q}, \lambda) = \int_{\mathbf{x}\in\Omega} \left(\mathbf{q}(\mathbf{x}) \cdot \delta\mathbf{q}(\mathbf{x}) + \lambda(\mathbf{x}) \ \nabla \cdot \delta\mathbf{q}(\mathbf{x}) \right) \mathrm{d}\mathbf{x} \tag{5.13}$$

with $\cdot$ denoting the inner product. Using partial integration, Eq. 5.13 can be rewritten as:

$$0 = \int_{\mathbf{x}\in\Omega} \left(\mathbf{q}(\mathbf{x}) - \nabla\lambda(\mathbf{x}) \right) \cdot \delta\mathbf{q}(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \int_{\mathbf{x}\in\partial\Omega} \lambda(\mathbf{x}) \left(\delta\mathbf{q}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) \right) \mathrm{d}\mathbf{x} \tag{5.14}$$

with $\mathbf{n}$ the inward normal to the boundary $\partial\Omega$ of Ω.

Typically, we either have that $\mathbf{q} \cdot \mathbf{n}$ on the boundary $\partial\Omega$ is known, and, thus, $\delta\mathbf{q} \cdot \mathbf{n} = 0$ on $\partial\Omega$ or that $\mathbf{q} \cdot \mathbf{n}$ on the boundary $\partial\Omega$ is unknown, which, because there is no constraint on $\mathbf{q}$ on the boundary, translates to $\lambda = 0$ on $\partial\Omega$. The latter, unconstrained boundary condition is referred to as free boundary in this paper, whereas the first, specified boundary condition has the employed closed boundary as a special example. With either $\delta\mathbf{q} \cdot \mathbf{n} = 0$ or $\lambda = 0$ on $\partial\Omega$, the last term of Eq. 5.14 equals zero, and Eq. 5.14 implies, since $\delta\mathbf{q}$ is arbitrary in the interior of Ω, $\mathbf{q}(\mathbf{x}) = \nabla\lambda(\mathbf{x})$. Therefore, we have

$$\mathbf{q}(\mathbf{x}) = \nabla\phi(\mathbf{x}) \tag{5.15}$$

with $\phi = \lambda$ satisfying Eq. 5.8 in the interior of Ω and either $\nabla\phi \cdot \mathbf{n} = 0$ or $\phi = 0$ on $\partial\Omega$.

This proves that the 2-norm of $\mathbf{q}$ is minimal if the vector field $\mathbf{q}$ is irrotational.

6 Scale-selective validation of morphodynamic models

This chapter is republished without noteworthy change from J. Bosboom and A. Reniers (2014). Scale-selective validation of morphodynamic models. In: *Proceedings 34th International Conference on Coastal Engineering*, Seoul, South-Korea, pp. 1911–1920, doi:10.9753/icce.v34.sediment.75.

Chapter 2 (Bosboom et al., 2014) and Ch. 3 (Bosboom and Reniers, 2018) identified the need for performance measures that quantify the agreement in patterns and features rather than the point-wise agreement as well as the need for performance measures that selectively address multiple spatial scales. The first point has led us to develop various error metrics that take the spatial structure of 2D morphological fields into account through a transformation of the computed towards the observed field, employing image warping (Bosboom and Reniers, 2014b, i.e. Ch. 4) and optimal transport (Bosboom et al., 2019, i.e. Ch. 5). In order to address the latter point, we developed a method that allows any metric to selectively address multiple scales. Our approach of employing a smoothing filter to compute localized statistics in a sliding window has a range of potential applications in our field, evident from the fact that it was taken up by Radermacher et al. (2018) with the aim of determining the sensitivity of modelled nearshore currents to errors in remotely-sensed bathymetries.

The highlights of Ch. 6 are:

1. A scale-selective validation approach is introduced that allows any metric to selectively address multiple spatial scales.
2. The method is relatively easy to implement and apply and yields results that are relatively straightforward to interpret.
3. Normalized measures for structural and amplitude simularity are introduced and combined into a pattern skill score.
4. The employed point-wise metrics do not suffer from the double penalty effect.
5. Areal maps of statistics are computed locally within a sliding window of progressively larger size.
6. A real-life application revealed strong spatial differences in structural and amplitude similarity and pattern skill and a lower prediction quality at the smaller scales.
7. The method can be used to determine the smallest scales with sufficient skill and establish the resolution at which model-data comparisons are ideally presented.

Abstract

Although it is generally acknowledged that the practical predictability at smaller scales may be limited, output of high-resolution morphodynamic area models is mostly presented at the resolution of the computational grid. The so-presented fields typically are realistic looking, but not necessarily of similar quality at all spatial scales. Unfortunately, commonly used single-number validation measures do not provide the necessary guidance as to which scales in the output can be considered skilful. Also, differences in skill throughout the model domain cannot be discerned. Here, we present a new, scale-selective validation method for 2D morphological predictions that provides information on the variation of model skill with spatial scale and within the model domain. The employed skill score weights how well the morphological structure and variability are simulated, while avoiding the double penalty effect by which point-wise accuracy metrics tend to reward the underestimation of variability. The method enables us to tailor model validation to the study objectives and scales of interest, establish the resolution at which results are ideally presented and target model development specifically at certain morphological scales.

6.1 Introduction

The traditional approach to morphodynamic model validation is to compute a single-number validation metric, such as the mean-squared error (MSE) or an MSE-based skill score (MSESS), for the entire 2D model domain or a limited number of subdomains (e.g. Sutherland et al., 2004). The validation of high-resolution morphodynamic models, however, brings about a range of new validation questions. Are there spatial displacement errors? Is the variability well represented at all scales? Is it necessary to accurately predict shorter-scale features to make reliable longer-term predictions? At which spatial scales does the model have sufficient skill? Does the skill vary within the model domain? These questions are not easily addressed with the traditional validation approach. Clearly, new techniques must be developed, which separately assess the various scales of interest in the morphology and patterns of bed change and take both similarity in structure and amplitude into account.

In other fields, notably meteorology, scale-dependent verification methods have been proposed that are able to describe the scale at which a forecast attains a particular level of skill (e.g. Roberts and Lean, 2008); for an overview, see Gilleland et al. (2010a). Also, in the field of image processing, Wang et al. (2003) determine the closeness of images using a multi-scale method, which incorporates image details at different resolutions. These methods typically utilize band-pass filters (Fourier, wavelets, etc.) or smoothing filters for the separation of scales. For 2D morphology and arbitrarily shaped model domains, the application of such band-pass filters and the physical interpretation of the results is far from trivial. Methods based on smoothing filters are appealing due to their simplicity, but often limited in the aspects of model performance that can be considered. For instance, no information on spatial variation of skill in the model domain is provided.

Fotheringham et al. (2002) analyze spatially varying relationships between measured variables by local regression modelling (i.e. in a neighbourhood around a regression point) and generalize this method to the computation of local weighted statistics in a sliding window. Our expectation is that such a conceptual framework, which allows the computation of a whole range of localized statistics, may not only be useful for data analysis but for model validation purposes as well.

The choice of validation metrics must be close to the intuitive judgement of morphologists. Point-wise accuracy metrics, such as the MSE, are useful, but tend to penalize, rather than reward, the model's capability to provide information on morphological features of interest (Bosboom and Reniers, 2014b; Bosboom et al., 2014, i.e. Ch. 4 and Ch. 2, respectively). Bosboom et al. (2014, Ch. 2) showed that this behaviour is also inherited by the MSESS and can be traced back to the implicit weighting in the MSE of the similarity in structure and amplitude of the fluctuations. To circumvent these issues, Taylor (2001) suggests an alternative weighting of these aspects.

In this paper, we present a new, scale-selective method for 2D morphological predictions that provides maps of prediction quality at various spatial scales. It bears similarities to localized data analysis (Fotheringham et al., 2002) in that it computes local validation metrics in a sliding window. The validation metrics are chosen to be close to the intuitive judgement of morphologists, viz. metrics pertaining to the structure and amplitude of the pattern and combined in a measure of pattern skill, in line with the skill score proposed by Taylor (2001). The various statistics are calculated for a range of window sizes, leading to maps of amplitude similarity, structural similarity and skill per scale. Note that the term "scale" is thus defined as geographical extent or areal size of focus. Aggregation of the results enables the determination of the smallest scale with useful domain-averaged skill. Attractive aspects of the method are the simplicity of implementation, application and interpretation of the results.

This paper is organized as follows: first our method of scaled skill is explained. Next, we demonstrate the method by comparing model predictions and data for the Bornrif, a dynamic attached bar at the Wadden Sea island of Ameland. Finally, the main conclusions are summarized.

6.2 Scaled skill

This section outlines our approach to quantify the skill and similarity in structure and amplitude per spatial scale as well as aggregated over all scales. First, we define normalized measures of amplitude and structural similarity and demonstrate that these can be expected to depend on the considered spatial scale, viz. geographical extent or areal size of focus. Next, we describe the method for deriving localized versions of these statistics. Finally, the approach is outlined to combine the maps of amplitude and structural similarity into a skill map per spatial scale and aggregate these maps for the entire model domain.

6.2.1 Aspects of model performance: structural and amplitude similarity per scale

A skilful model should be able to accurately simulate both the structure and the variance of fluctuating signals. These notions can be represented by the correlation ρ_{po} and the ratio of the standard deviations of predictions and observations $\hat{\sigma} = \sigma_p/\sigma_o$ (Bosboom et al., 2014, i.e. Ch. 2). The correlation ρ_{po} (with $-1 \leqslant \rho_{po} \leqslant 1$) measures the tendency of observations and predictions to vary together. A non-perfect correlation, i.e. smaller than unity, may result from incorrect locations, shapes and *relative* magnitudes of features. A value of $\hat{\sigma} = \sigma_p/\sigma_o$ larger or smaller than 1 indicates an overestimation or underestimation, respectively, of the variance of the signal.

In the following, we use the correlation as a normalized measure of the structural similarity between predictions and observations. We further define a normalized measure for amplitude similarity:

$$\eta = \left(\frac{2}{\hat{\sigma} + \hat{\sigma}^{-1}}\right)^q, \quad 0 \leqslant \eta \leqslant 1 \tag{6.1}$$

with q a coefficient (set to 2 in this paper). Perfect agreement is indicated by $\eta = 1$. As opposed to $\hat{\sigma}$, the parameter η is bounded and invariant under the exchange of predictions and observations. Hence, overprediction and underprediction are now equally penalized. When it is important to distinguish between over- and underprediction, $\hat{\sigma}$ can be used. Note that Eq. 6.1 can be rewritten as:

$$\eta = \left(\frac{2\sigma_p\sigma_o}{\sigma_p^2 + \sigma_o^2}\right)^q \tag{6.2}$$

which, with $q = 1$, is the form as used by Wang et al. (2003) and Koh et al. (2012) and named contrast measure and variance similarity, respectively.

In morphodynamic modelling, where the predictand is the bathymetry, the interpretation of ρ_{po} and $\hat{\sigma} = \sigma_p/\sigma_o$ in terms of bed features is far from trivial, since multiple scales are generally present in the observed and computed bathymetry (Fig. 6.1) and larger scales may overwhelm the smaller scales. Figure 6.1 (middle panel) indicates that the overall correlation can be negative, whilst the correlation can be positive if we zoom in to a smaller area. This situation can of course also be reversed, with positive correlation for larger scales and negative correlation for smaller scales (Fig. 6.1, top panel). The latter situation may be closer to what we expect from a typical morphodynamic simulation. Not only the correlation but also the ratio of the standard deviations between predictions and observations may vary with spatial scale. For example, Fig. 6.1 (bottom panel) shows an overestimation of the variability for the larger scale and an underestimation for the smaller scale.

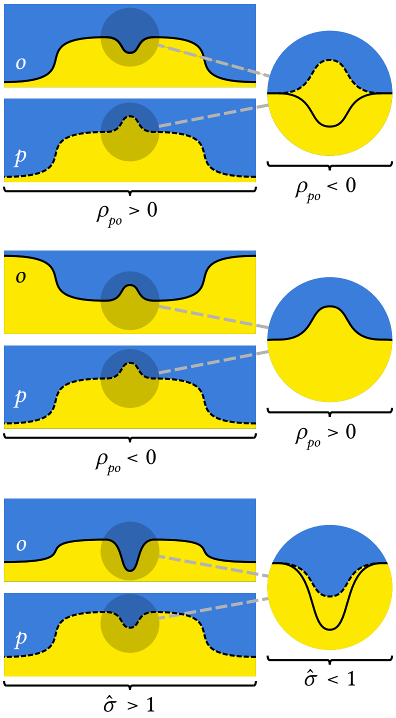


Figure 6.1: Scale-dependency of comparisons between observations o and predictions p. *Top panel*: the correlation is higher at the larger scale; *middle panel*: the correlation is higher at the smaller scale; and *bottom panel*: the amplitude similarity is also dependent on the scale.

6.2.2 Localized statistics

In order to generate maps of localized statistics, the structural and amplitude similarity are computed locally within a sliding window that moves across the domain. Herewith, we obtain fields of localized statistics for a particular window size. In order to account for various spatial scales, viz. areas of different geographical extent, we repeat this process for multiple window sizes.

For the ith grid point the local weighted means $\overline{o}_i$ and $\overline{p}_i$ of the observations o and predictions p, respectively, are given by:

$$\bar{o}_i = \sum\nolimits_j w_{ij} o_j \tag{6.3}$$

$$\bar{p}_i = \sum\nolimits_j w_{ij} p_j \tag{6.4}$$

with w_{ij} is a weighting factor dependent on the proximity to the location i and $\sum_j w_{ij} = 1$. All results shown in this paper are obtained with a very simple (and fast) window, viz. a rectangular window with a width W, uniform weights within the window and $w_{ij} = 0$ elsewhere in the domain (Fig. 6.2). Hence, $w_{ij} = w_{ij}(W)$. A more sophisticated approach uses a distance decay function given by for instance a bi-square kernel with a variable bandwidth (see e.g. Fotheringham et al., 2002).

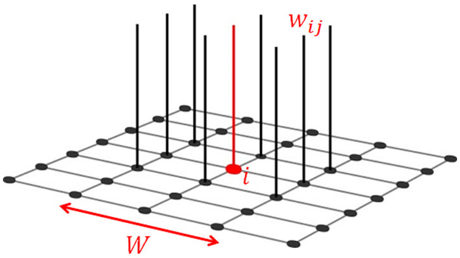


Figure 6.2: The rectangular window, around grid point i, with window width W and weights w_{ij}.

Of course, Eqs. 6.3 and 6.4 simply compute a (weighted) moving average. However, we can now extend the concept to arbitrary statistics, for instance the standard deviations $\sigma_{o,i}$ and $\sigma_{p,i}$ of observations and predictions, respectively:

$$\sigma_{o,i} = \left[\sum\nolimits_j w_{ij} \left(o_j - \bar{o}_i\right)^2\right]^{1/2} \tag{6.5}$$

$$\sigma_{p,i} = \left[\sum\nolimits_j w_{ij} \left(p_j - \bar{p}_i\right)^2\right]^{1/2}. \tag{6.6}$$

Similarly, the local correlation $\rho_{po,i}$ between predictions and observations is determined by:

$$\rho_{po,i} = \left(\sigma_{o,i}\sigma_{p,i}\right)^{-1} \sum\nolimits_j w_{ij} \left(o_j - \bar{o}_i\right)\left(p_j - \bar{p}_i\right), \quad -1 \leqslant \rho_{po,i} \leqslant 1. \tag{6.7}$$

Note that in Eqs. 6.6 and 6.7, the local rather than the global mean values are used. Now, the local amplitude similarity is given by:

$$\eta_i = \left(\frac{2}{\hat{\sigma}_i + \hat{\sigma}_i^{-1}}\right)^q \quad \text{with} \;\; \hat{\sigma}_i = \sigma_{p,i}/\sigma_{o,i}, \quad 0 \leqslant \eta_i \leqslant 1. \tag{6.8}$$

Note that all above statistics, which are formulated in terms of bed levels, could also be formulated in terms of cumulative bed change.

6.2.3 How to construct a skill score?

The correlation between predictions and observations and the ratio of the standard deviations of predictions and observations are important ingredients of the often used accuracy measure MSE. The fluctuating or pattern part of the MSE can be written as (see e.g. Bosboom et al., 2014, i.e. Ch. 2):

$$\mathrm{MSE}_{\mathrm{fluct}} = \sigma_o^2 \left[1 - \rho_{po}^2 + \left(\rho_{po} - \hat{\sigma} \right)^2 \right]. \tag{6.9}$$

Between two predictions with the same positive correlation, $\mathrm{MSE}_{\mathrm{fluct}}$ is minimized for $\hat{\sigma} = \rho_{po}$, hence for $\sigma_p = \rho_{po}\sigma_o$. In the case of a negative correlation, $\mathrm{MSE}_{\mathrm{fluct}}$ is minimized for $\hat{\sigma} = 0$ and thus for $\sigma_p = 0$. As a consequence, the MSE tends to reward the underestimation of the variability (Bosboom et al., 2014, i.e. Ch. 2).

Nonetheless, a morphologist may prefer features to be predicted at the right amplitude albeit displaced above a featureless prediction (Bosboom and Reniers, 2014b, i.e. Ch. 4). Therefore, we use an alternative weighting with the following behaviour: for any given variance, the skill score increases monotonically with increasing correlation and for any given correlation the skill score increases as the modelled variance approaches the observed variance (Taylor, 2001).

A general form for a local pattern skill score in terms of the normalized measures for structural similarity $\rho_{po,i}$ and amplitude similarity η_i then reads:

$$S_i = \frac{1}{2}(1 + \rho_{po,i})^m \eta_i^n, \quad 0 \leqslant S_i \leqslant 1. \tag{6.10}$$

Note that S_i is a function of the window width W. The weighting of structural and amplitude similarity must, to a certain extent, be decided upon subjectively. The coefficients m and n allow the user to define the most appropriate weighting for the situation under consideration. In this paper, we have used $m = 1$ and $n = 1$ in Eq. 6.10 and $q = 2$ in Eq. 6.8. A domain-averaged skill score S as a function of W can be obtained by averaging S_i (Eq. 6.10) over all grid points i.

We hypothesize that the smaller scales, down to the grid scale, are not as well predicted as the larger scales up to the scale of the entire domain, and that there is a minimum spatial scale above which the skill is sufficient, i.e. larger than a user-defined target skill (Fig. 6.3). For a real-life case, this hypothesis is put to the test in the next section.

6.3 Example

In this section, we demonstrate our method by applying it to measured and computed bathymetric fields for the Bornrif, a dynamic attached bar at the Northwestern edge of the Wadden Sea island of Ameland. First, we briefly describe the measurements and computations. Next, we show the maps of local statistics,

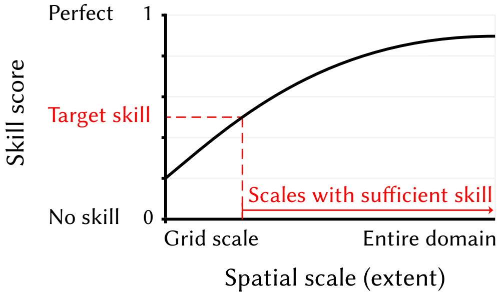


Figure 6.3: Hypothesized qualitative behaviour of the skill score S versus the spatial scale, which ranges from the grid scale to the entire domain. For larger spatial scales the skill value approaches the whole-map skill value $S = 1/2(1 + \rho_{po})\eta$ computed using the values at all grid points.

which are subsequently pooled into map-mean values per spatial scale. Finally, we explore the relationship between information richness and skill.

6.3.1 Bornrif

The Bornrif morphodynamic evolution was computed with Delft 3D from 1993 to 2008, using a grid with a resolution of $50 \times 50\,\mathrm{m}^2$ in the central part of the model domain and $100 \times 50\,\mathrm{m}^2$ closer to the model boundaries (Achete et al., 2011). A detailed description of this Delft3D simulation and the available data is found in (Bosboom et al., 2014, i.e. Ch. 2). Here, we focus on the results for 1998, hence five years after the start of the simulation (Fig. 6.4).

Upon visual comparison of the 1998 computations and data, we can observe differences at various locations and spatial scales. For instance, note the differences in the position and extent of the overall shape of the Bornrif as well as of the spit that has just attached to the mainland. Further, at relatively large water depths to the east of the Bornrif, sand bars are clearly visible in the observations, but largely absent in the computations. The area closest to the inlet, to the west of the Bornrif is characterized by multiple channels that are not well represented in the computations. Also of interest are the nearshore regions; east of the Bornrif, the measurements show multiple bars, which are not reproduced by the model. Further, differences can be observed in the slopes of the relatively steep near-shore regions, especially along the west flank of the Bornrif, which are crucial for the magnitude of the alongshore transport.

The analysis region, as shown in Fig. 6.4, covers only that part of the computational domain for which data are available during the entire simulation duration. In order to retain all observed scales, the spatial validation analysis is performed on the $20 \times 20\,\mathrm{m}^2$ grid that the data were presented on. To that end, the computations were first interpolated onto the observational grid. In the following we demonstrate typical results of applying the method of scaled skill. The central

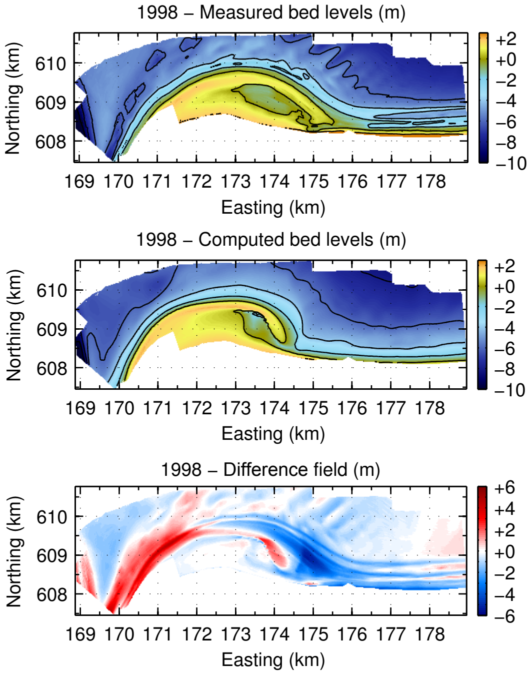


Figure 6.4: Measured (*top panel*) and computed (*middle panel*) Bornrif bathymetries for 1998 and the difference field $p - o$ between predictions p and observations o (*lower panel*).

validation question is: how skilful is the model in the various regions and at the various spatial scales that can be discerned?

6.3.2 Maps of local statistics

Areal maps of structural similarity $\rho_{po,i}$, amplitude similarity $\hat{\sigma}_i$ and η_i, and pattern skill S_i provide information on local differences in quality (Fig. 6.5). Such maps can be produced for various spatial scales (i.e. areal sizes of focus). Figure 6.5 shows the results at three window sizes W = 0.16, 0.4 and 0.8 km. There is a wealth of information in these figures; here we will only point out some main aspects.

The negative correlation in the area west of the Bornrif clearly indicates the lack of structural similarity between the two patterns, except close to the coastline where the correlation is higher again. This dissimilarity is quite persistent as the spatial scale increases. Another patch with negative correlations at all scales is the result of the computed spit being present at the observed lagoon. In the spit area, the largest dissimilarity in amplitude is found somewhat further offshore, reflecting the fact that the computed slope is clearly off.

On the contrary, there are also small-scale patches of negative correlation that are not present anymore at the larger scales, for instance in regions further offshore and in the nearshore region east of the Bornrif. In these areas, a low struc-

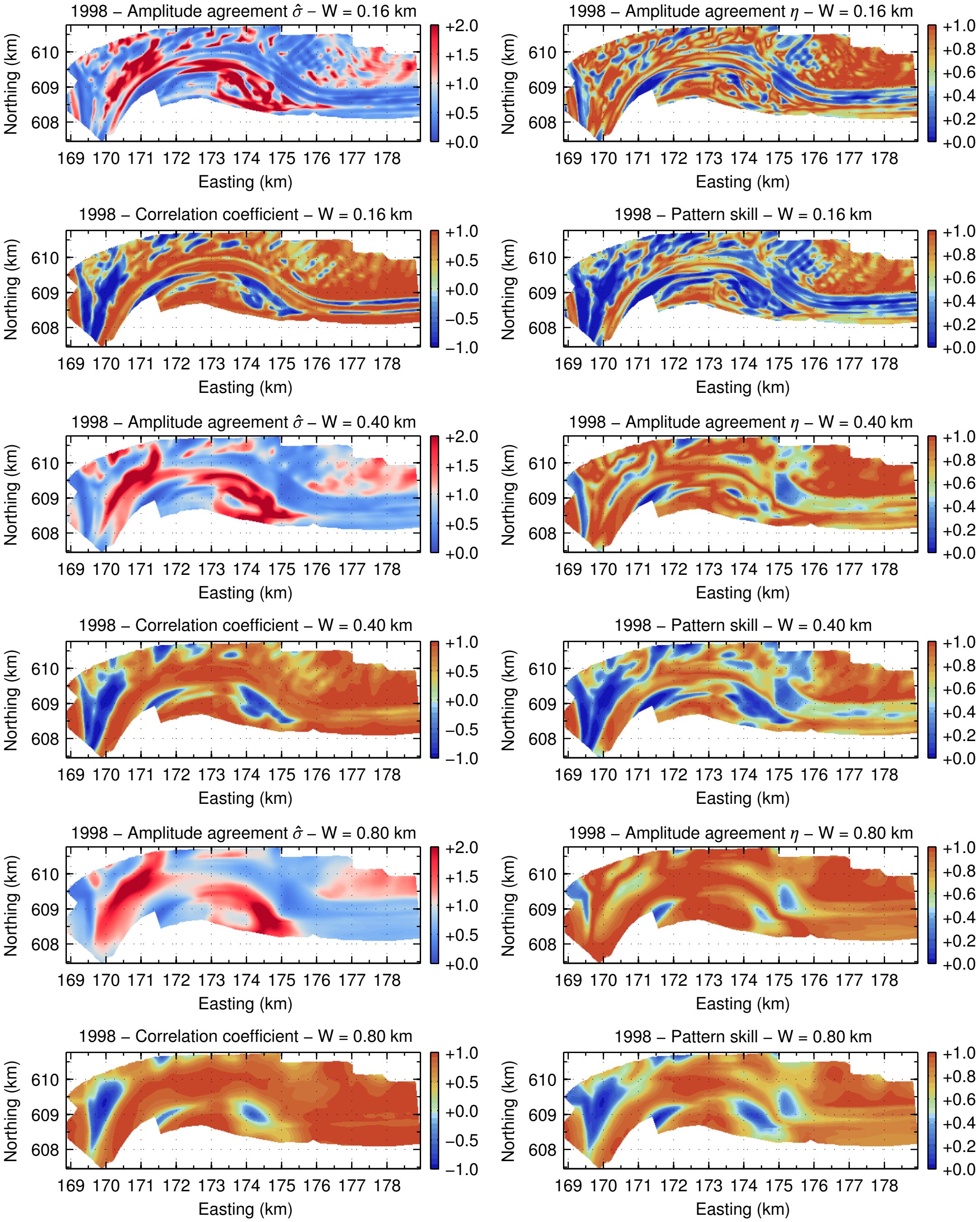

1998 – Amplitude agreement $\hat{\sigma}$ – W = 0.16 km
1998 – Amplitude agreement η – W = 0.16 km
1998 – Correlation coefficient – W = 0.16 km
1998 – Pattern skill – W = 0.16 km
1998 – Amplitude agreement $\hat{\sigma}$ – W = 0.40 km
1998 – Amplitude agreement η – W = 0.40 km
1998 – Correlation coefficient – W = 0.40 km
1998 – Pattern skill – W = 0.40 km
1998 – Amplitude agreement $\hat{\sigma}$ – W = 0.80 km
1998 – Amplitude agreement η – W = 0.80 km
1998 – Correlation coefficient – W = 0.80 km
1998 – Pattern skill – W = 0.80 km
Northing (km)
610
609
608
169 170 171 172 173 174 175 176 177 178
Easting (km)
+2.0
+1.5
+1.0
+0.5
+0.0
+1.0
+0.8
+0.6
+0.4
+0.2
+0.0
+1.0
+0.5
+0.0
−0.5
−1.0

Figure 6.5: Normalized maps of structural and amplitude similarity and pattern skill for three different window sizes W = 0.16, 0.4 and 0.8 km. For all quality metrics a value of 1 represents perfect agreement.

tural similarity $\rho_{po,i}$ is combined with a low amplitude similarity η_i, which can be seen—from $\hat{\sigma}_i$ being close to zero—to be due to an underestimation of the variability. This indicates small-scale, observed features that are not reproduced in the predictions, namely the sand bars at deeper water and the nearshore bars.

As expected, the maps of pattern skill can be seen to combine the characteristics of the maps of structural and amplitude similarity. At the smallest window width, the skill areal maps show relatively large areas with low skill. At larger window widths only the larger-scale deviations remain.

6.3.3 Pooled skill scores

Another way of looking at the quality variation is by making histograms of the quality maps (Fig. 6.6). The first column clearly shows that grid points with negative correlation at small spatial scales obtain a positive correlation at larger scales. A similar trend can be observed from the second column that shows the amplitude similarity. The third column shows that, as a result, the percentage of the model domain with low pattern skill scores decreases with spatial scale, as was apparent from the pattern skill maps as well (Fig. 6.5).

The red lines in Fig. 6.6 show the domain-averaged values of the quality metrics for the three window sizes that are considered. Not surprisingly given the above, the quality according to each of these metrics increases with spatial scale. Apparently, the Bornrif morphology can be thought to consist of smaller-scale features that are not well represented by the model, on top of a larger-scale morphology that is better predicted.

When extending this analysis to a range of window sizes, we obtain Fig. 6.7, which shows the structural similarity, amplitude similarity and pattern skill versus window size. At the scale of the entire domain, the skill is very high, since the larger-scale morphology is reasonably well represented. However, at the smaller scales of the spit and the sand bars the skill is lower. Based on this figure, we can determine the smallest useful scale, viz. the smallest areal size with a certain desired level of skill. If the target skill (see Fig. 6.3) is set to for instance 0.7, the smallest scale with sufficient skill is about 0.6 km.

6.3.4 Information content versus skill

Output of high-resolution morphodynamic area models is generally presented at the resolution of the computational grid. The previous findings suggest, however, that the high-resolution detail may not be skilful. Consequently, a smoother bathymetry (Fig. 6.8) may be more skilful than the original, computed bathymetry (Fig. 6.4). The bathymetries in the left and right columns of Fig. 6.8 are obtained by applying a moving average to the original bathymetries, using window sizes of $W = 0.4$ and 1.6 km, respectively (using Eqs. 6.3 and 6.4).

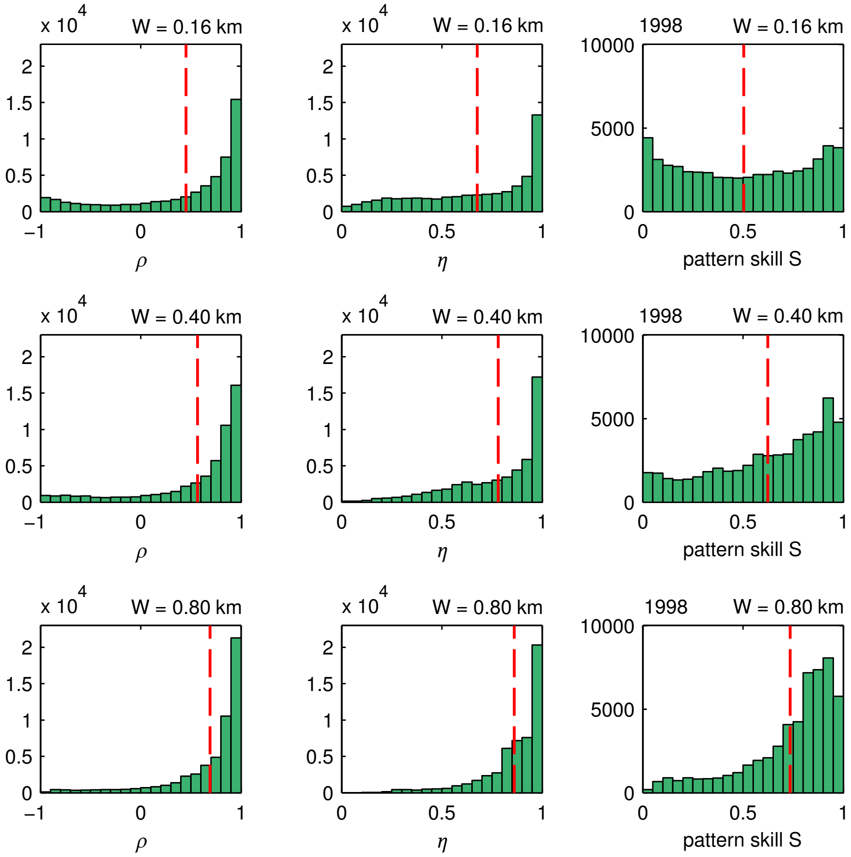


Figure 6.6: Histograms of the correlation, amplitude similarity and pattern skill for the three window sizes W = 0.16, 0.4 and 0.8 km. Note that the histograms correspond to the respective maps in Fig. 6.5. The red lines indicate the domain-averaged values which can be seen to increase with spatial scale.

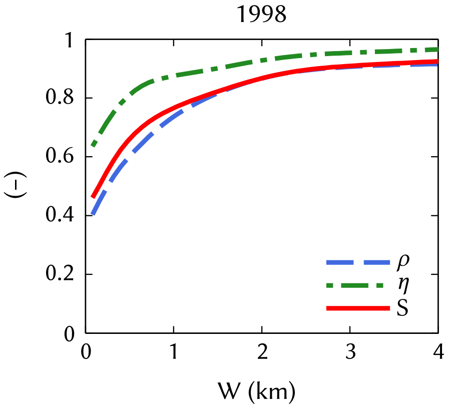


Figure 6.7: Structural and amplitude similarity and pattern skill as a function of window size.

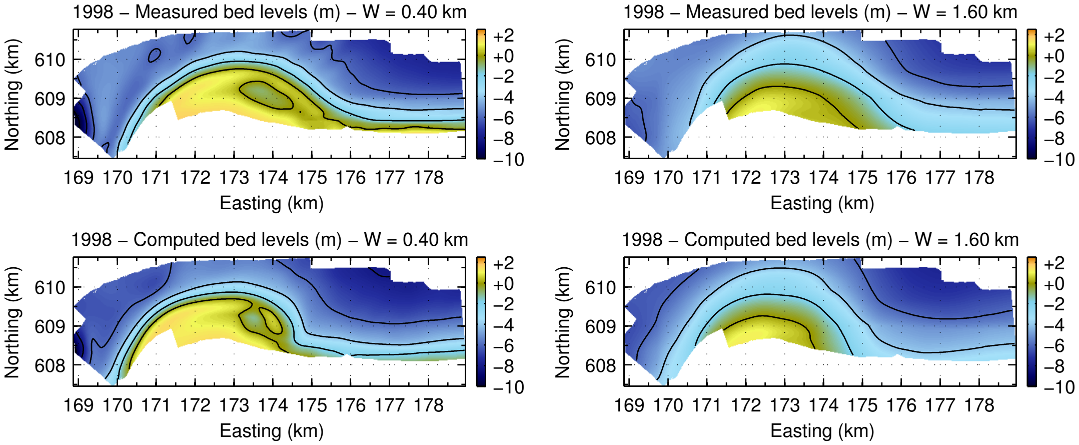

1998 – Measured bed levels (m) – W = 0.40 km
1998 – Measured bed levels (m) – W = 1.60 km
1998 – Computed bed levels (m) – W = 0.40 km
1998 – Computed bed levels (m) – W = 1.60 km
Northing (km)
608
609
610
Easting (km)
169 170 171 172 173 174 175 176 177 178
+2
+0
−2
−4
−6
−8
−10

Figure 6.8: Spatial means, obtained by Eqs. 6.3 and 6.4, of the original high-resolution bathymetries. *Left*: W = 0.4 km and *right*: W =1.6 km.

To determine the effect of leaving the high-resolution detail out, we apply the same validation procedure as before, at a range of window sizes, but now not to the full-resolution bathymetries, but to their smoothed counterparts. The aggregated results are shown in Fig. 6.9. For clarity, the skill trend for the full-resolution bathymetries (Fig. 6.7) is repeated in Fig. 6.9. The latter figure confirms that for all scales the presented smoother bathymetries are more skilful. Note that for the bathymetries smoothed with $W = 0.4$ km, all scales have a skill around or above the target skill of 0.7.

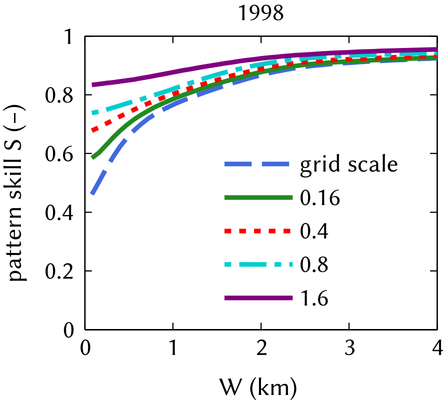


Figure 6.9: Pattern skill versus window size for bathymetries with a different level of smoothening (a moving average at window sizes ranging from 0.16 km to 1.6 km. The pattern skill at full resolution (Fig. 6.7) is repeated here and indicated with “grid scale”.

Evidently, the inclusion of smaller scales, up to the full model resolution, contributes negatively to the skill at especially the smaller scales. Of course, the increase in skill for smoother bathymetries comes at a loss of information richness; the smoothed bathymetries are less realistic looking than the full resolution bathymetries. Ideally, the computational results should be presented at a scale that finds a balance between skill and information richness.

6.4 Conclusions

We have presented a scale-selective validation method for 2D morphological predictions that allows the computation of localized statistics at various spatial scales and the generation of areal maps of these statistics. The term “scale” refers to geographic extent or areal size of focus. In this paper, we use normalized measures of structural and amplitude similarity and combine these in a measure of morphological pattern skill, but other validation metrics can be used as well. Also, the method could be supplemented with a bias term at the largest scale.

Application to the Bornrif showed strong spatial differences in structural and amplitude similarity and pattern skill. Further, due to amongst others small-scale observed features that are not (well) reproduced in the predictions, a lower domain-averaged prediction quality was found at the smaller scales than at the larger scales. In relation to this, it was found that smoothing out the high-resolution detail increases the skill of the results especially at the smaller scales, even though the smoothed bathymetries are less realistic looking than the full-resolution bathymetries.

In summary, the method can be used to:

1. Determine local differences in structural and amplitude similarity and pattern skill;
2. Determine the smallest scales with sufficient skill;
3. Establish the resolution at which model-data comparisons are ideally presented;
4. Target model development specifically at certain morphological scales.

Compared to possible alternative strategies to scale-selective model validation, the method is easy to implement and apply, and the results are relatively easy to interpret. This makes it a tool that can be readily used for practical purposes.

7 Conclusions and recommendations

The overarching aim of this thesis was to contribute to an improved validation assessment of morphological predictions, in particular field predictions (Sect. 1.3). To that end, Chs. 2 to 6 pursued two main research objectives, which were derived from Sects. 1.1 and 1.2, respectively. Conclusions pertaining to these two objectives are discussed in Sects. 7.1 and 7.2, respectively. Sect. 7.3 discusses implications for model validation as well as recommendations for further research.

7.1 The behaviour of the $\mathrm{MSESS_{ini}}$ a.k.a. the BSS

The first main research objective (Objective 1) was to investigate the behaviour of the commonly used mean-squared-error skill score $\mathrm{MSESS_{ini}}$ a.k.a. the Brier skill score (BSS)[1] with the initial bed as the reference prediction. It was elaborated in four research questions (1.1 to 1.4, see Sect. 1.3) and addressed in Ch. 2 (Bosboom et al., 2014) and Ch. 3 (Bosboom and Reniers, 2018). The main findings pertaining to each of these four research questions are summarized in Sects. 7.1.1 to 7.1.4.

7.1.1 Inheritance from the MSE: smooth is better

Question 1.1 was formulated as follows: what is the effect on the $\mathrm{MSESS_{ini}}$ of the use of the point-wise mean-squared error (MSE) as the accuracy measure? This research question was answered in Chs. 2 and 3 (Bosboom et al., 2014; Bosboom and Reniers, 2018).

In the presence of inevitable location errors, the MSE and other overall point-wise metrics are prone to penalize rather than reward the correct prediction of variability (Anthes, 1983; Arpe et al., 1985; Taylor, 2001). As a consequence, featureless predictions are sometimes favoured over predictions whose features are misplaced, a characteristic that is referred to as “double penalty effect” (Bougeault, 2003). In Chs. 2 and 3, it was demonstrated that the double penalty effect is inherited by the MSESS a.k.a. BSS, resulting in a tendency to reward the underestimation of the variance of morphodynamic change relative to the reference prediction. In the case of $\mathrm{MSESS_{ini}}$, hence using the initial bed as the reference

[1] This thesis addresses the MSE-based skill metric for nonprobabilistic variables as mean-squared-error skill score (MSESS), consistent with Murphy (1988). The subscript “ini” specifies that the reference prediction used is the initial bed at the start of the simulation. In our field, the $\mathrm{MSESS_{ini}}$ is known as the Brier skill score (BSS). Technically however, the term Brier skill score is reserved for the relative accuracy of probabilistic forecasts with the Brier score (Brier, 1950) as the accuracy measure, which is a mean-squared error for probabilistic forecasts with two mutually-exclusive outcomes (e.g. rain or no rain).

prediction, this implies an underestimation of the overall magnitude of the cumulative bed changes from the start of the simulation. As a consequence, predictions of sedimentation/erosion features that are correct in terms of magnitude but are misplaced in space may not outperform even the reference prediction of zero change, as was nicely illustrated by a numerical hindcast of morphological changes of a wide estuary mouth sandbank, located along the French Atlantic Coast (Guerin et al., 2016)—out of two morphodynamic simulations, the simulation that captures several of the main morphological changes receives a lower score ($\text{MSESS}_{\text{ini}} = -0.18$) than the simulation that predicted almost no morphological change ($\text{MSESS}_{\text{ini}} = 0.01$).

The tendency of the $\text{MSESS}_{\text{ini}}$ to reward the underestimation of the variance of bed changes was demonstrated through the behaviour of the amplitude error or, rather, scale error β', which follows from the Murphy–Epstein decomposition of the MSE-based skill score (Sect. 7.1.2). The scale error β' depends on the correlation $\rho_{p'o'}$ between the predicted (p') and observed (o') cumulative bed changes relative to the reference prediction, and on $\sigma_{p'}/\sigma_{o'}$, the ratio of the predicted over the observed standard deviation of the bed changes. For positive, suboptimal anomaly correlation ($0 < \rho_{p'o'} < 1$), $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'}$ minimizes β' and—unless compensated by systematic bias—maximizes the skill $\text{MSESS}_{\text{ini}}$, whereas for $\rho_{p'o'} \leqslant 0$, $\sigma_{p'} = 0$ maximizes the skill. In other words, for the same suboptimal anomaly correlation and systematic bias, a higher skill value is found for sedimentation/erosion fields that underpredict the overall amount of sedimentation and erosion than for predictions with the correct variance of the bed changes. This runs contrary to the morphologists' intuition of optimal performance requiring that $\sigma_{p'} = \sigma_{o'}$.

Clearly, these findings have implications for (automated) calibration procedures that optimize the MSE or $\text{MSESS}_{\text{ini}}$ (e.g. Briere et al., 2011; Simmons et al., 2017). In modelling practice, a reduction of the overall size of bed changes is easier to achieve, for instance by changing the grain size or a transport parameter, than an improvement of the anomaly correlation coefficient. As a consequence, the reduction of the overall sizes of bed changes is an effective, though undesirable method to obtain higher values of $\text{MSESS}_{\text{ini}}$. This was illustrated by a real-life example, taken from Sutherland et al. (2004), of the comparison of observed bathymetric changes for East Pole Sand with three field predictions. The predictions only differ with respect to the representative grain-size parameter and yield the same positive, but nonperfect anomaly correlation coefficient. As a consequence, the largest $\text{MSESS}_{\text{ini}}$ (and the smallest MSE) are achieved by the prediction with the coarsest grain size, which shows—with $\sigma_{p'}/\sigma_{o'} \approx \rho_{p'o'}$—the most severe underprediction of the variance of the bed changes. Another example was provided by the morphodynamic simulations of the Bornrif (Achete et al., 2011), a dynamic attached bar at the Wadden Sea island of Ameland, which was analyzed in depth in Bosboom et al. (2014), i.e. Ch. 2. Throughout the 15 years of simulation, $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$ are close together in value with in the last year (2008) $\rho_{p'o'} \approx \sigma_{p'}/\sigma_{o'} \approx 0.66$

and $\mathrm{MSESS_{ini}} = 0.45$. For $\sigma_{p'}/\sigma_{o'} = 1$, however, the skill would have been lower at $\mathrm{MSESS_{ini}} = 0.33$.

It may therefore well be that in many modelling studies, without the modeller necessarily being aware of this, the ratio of predicted over observed anomaly standard deviation is lowered towards the level of the correlation. Although this certainly optimizes the MSE and $\mathrm{MSESS_{ini}}$, another aspect of model quality, the variance of bed changes, is less well predicted. We therefore advocate that values of $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$ are also explicitly reported and a deliberate choice is made for the optimal simulation.

7.1.2 Additional insight through the Murphy–Epstein decomposition

Question 1.2 was formulated as follows: what is the added value and correct interpretation of the Murphy–Epstein decomposition of the $\mathrm{MSESS_{ini}}$? This research question was answered in Chs. 2 and 3 (Bosboom et al., 2014; Bosboom and Reniers, 2018).

An advantage of the mean-squared-error measure of accuracy and the corresponding MSESS is that they can readily be decomposed into components that describe specific elements of prediction quality. The decomposition according to Murphy and Epstein (1989) separates the MSE, either expressed in terms of bed levels or in terms of bed changes relative to the reference (the anomalies), into correlation and conditional bias terms, which quantify the mismatch between the fluctuating parts in predictions and observations, and systematic bias or map-mean error (Appendix 2.A). Using the MSE decomposition in terms of the anomalies, the MSESS can be decomposed into a correlation term or phase error $1 - \alpha'$, conditional bias or amplitude error β' and normalized systematic biases or map-mean errors of the prediction γ' and the reference predition ϵ' (Eqs. 2.7, 2.9 and 2.10). Although this decomposition can provide valuable insight into specific aspects of prediction quality, it has only been used in a limited number of morphological applications and seems to be not well understood (see Chs. 2 and 3). In Chs. 2 and 3, we have investigated what exactly is measured by the separate error terms, using the real-life application of the Bornrif (Achete et al., 2011) and a series of simple, hypothetical test cases, respectively.

Chapter 3 demonstrated that α' and β' are best explained from the linear least-squares regression of o' given p' and are preferably referred to as structural similarity and scale error, respectively. The term α' is the square of the correlation and indicates the tendency of o' and p' to vary together and, hence, expresses the similarity in the structure of the sedimentation/erosion fields. To establish the *direction* of the linear correspondence $\rho_{p'o'}$ must be evaluated instead of α'. A zero value for β', which is the squared difference between $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$, indicates an optimal scaling of the magnitude of the anomalies to account for a nonperfect, nonnegative value of the correlation ($0 \leqslant \rho_{p'o'} < 1$). "Optimal" is defined here in terms of the smallest overall least-squares error and skill score and not in terms of the variance of the bed changes, which must be judged separately and requires that $\rho_{p'o'}$, $\sigma_{p'}/\sigma_{o'}$ and β' are reported (see Sect. 7.1.1).

For the interpretation of γ' and ϵ' for morphological models, it is important to realize that γ' is a sediment budget error of the prediction and ϵ' of the reference prediction, often a zero change prediction, and that both error terms are normalized by the variance of the observed cumulative change away from the reference prediction. Hence, in practice, we expect that both terms are generally small, especially towards the end of a simulation. For the Bornrif, for instance, γ' and ϵ' were found to be small and decrease towards zero during the 15 years of simulation. Since also β' is small throughout the simulation, with in the last year (2008) $\beta' \approx 0$, the skill in 2008 is equal to the so-called potential skill in the absence of biases; $\text{MSESS}_{\text{ini}} \approx \alpha' = \rho^2_{p'o'} = 0.45$ (see Sect. 7.1.1). Similarly, the "optimal" prediction for East Pole Sand yielded β', γ' and ϵ' close to zero and, consequently, $\text{MSESS}_{\text{ini}}$ close to $\rho^2_{p'o'}$.

These calibration examples may well be representative for many morphodynamic model studies, elucidating that if $\text{MSESS}_{\text{ini}}$ is used, it is best supplemented with $\rho_{p'o'}$, $\sigma_{p'}/\sigma_{o'}$, α', β', γ' and ϵ' for a good interpretation of model performance.

7.1.3 How to correctly account for measurement error?

Question 1.3 was formulated as follows: what is the rationale behind taking measurement error into account and how should this translate to skill formulations and rankings? This research question was answered in Ch. 3 (Bosboom and Reniers, 2018).

Generally, MSE-based skill formulations assume that the MSE of a perfect prediction is zero, which means that the presence of errors in the data is not taken into account. In reality, data errors will occur due to errors in the bathymetric surveys as well as in the subsequent interpolation procedure to a common grid. As a consequence, a deviation from 1 of the value for $\text{MSESS}_{\text{ini}}$ is not purely due to prediction error, but can partly be attributed to data errors. This thesis has evaluated the adjustments to the MSESS as proposed by Sutherland et al. (2004) and van Rijn et al. (2003) with the aim to account for measurement error (Sect. 3.4) as well as their proposed classifications of prediction quality, for the skill scores with and without corrections for measurement error (Sect. 3.5). This has led to the conclusion that the existing methods to correct for measurement error are inconsistent in either their skill formulation or their suggested classification scheme, which will be briefly explained below.

The adjusted skill formulation by Sutherland et al. (2004), i.e. Eq. 3.23, is based on the assumptions that the initial and final measured bathymetries consist of an actual bathymetry and independent, random measurement errors δ with the same $\delta_{\text{rms}} = \sqrt{\langle \delta^2 \rangle}$ and that the initial error persists throughout the simulation.

Consequently, perfect skill is already obtained for MSE $= 2\langle\delta^2\rangle$ instead of for MSE $= 0$ and positive skill values increase. We demonstrated that the adjusted skill formulation by Sutherland et al. (2004) is equivalent to the original skill formulation applied to that part of the mean-squared errors that can be attributed to the predictions. Hence, as long as the actual errors of the reference prediction and prediction remain unchanged, also the skill values remain unchanged. The adjusted formulation of van Rijn et al. (2003), i.e. Eq. 3.25, was shown to partly adjust for measurement error in the MSE of the prediction, but to fail to correct for its effect on the MSE of the reference prediction. As a consequence, the skill values increase in the presence of measurement error, even if the actual errors of the reference prediction and prediction remain unchanged. Therefore, as already advocated by Sutherland et al. (2004), Eq. 3.23 is the recommended formulation to take measurement error into account, since, as opposed to Eq. 3.25 by van Rijn et al. (2003), it is consistent with the definition of skill.

Both Sutherland et al. (2004) and van Rijn et al. (2003) propose an alternative skill classification scheme, to be used in conjunction with their adjusted skill formulation. We have argued, however, that a skill formulation that is effective in removing measurement error yields values that can and should be directly compared to the classification valid in the absence of data errors.

In order to further substantiate this claim, we designed an artificial case of the formation of two rip channels, slightly different in shape and position. A planar beach served as the initial bathymetry, while the beach modified with the one rip channel served as the observations and with the other as the predictions. By adding noise fields δ to these bathymetries, we could compute the skill values as a function of measurement error δ_{rms} according to the original skill formulation as well as the two previously described adjusted formulations. As expected, skill scores according to the original skill formulation decrease with increasing δ_{rms}, such that the predictions are unjustly penalized for the presence of measurement error. The addition of measurement error, regardless of the magnitude of δ_{rms}, does not change the skill scores according to the formulation of Sutherland et al. (2004), which, thus, proves to be effective in removing the influence of the measurement error. The constant skill irrespective of δ_{rms} confirms, however, that an adjusted classification of the skill score is not appropriate. Van Rijn's adjusted formulation (van Rijn et al., 2003), on the contrary, gives a strong increase of the skill scores with δ_{rms}. We expect that the suggestion to adjust the skill ranking stems from this inflation of the skill scores when adding measurement error. A universal adjusted ranking, however, cannot exist since the required adjustment depends on the measurement error and the prediction situation.

7.1.4 The ranking based on $\mathrm{MSESS}_{\mathrm{ini}}$ a.k.a. the BSS is not generally valid

Question 1.4 was formulated as follows: to what extent does the zero change model underlying the $\mathrm{MSESS}_{\mathrm{ini}}$ make model performance comparable across different prediction situations—geographical locations, forcing conditions, time periods, internal dynamics? This research question was answered in Chs. 2 and 3 (Bosboom et al., 2014; Bosboom and Reniers, 2018).

Since the $\mathrm{MSESS}_{\mathrm{ini}}$ uses the initial observed bathymetry at the start of a simulation as the reference, the zero point at the scale of morphodynamic model skill is set by a model that predicts zero morphological change. Therefore, a comparative analysis based on $\mathrm{MSESS}_{\mathrm{ini}}$ can only be effective if the intrinsic difficulty of common prediction situations is correctly reflected in the level of accuracy of the zero change model. In Chs. 2 and 3, it is demonstrated through various simple and real-life examples that the zero change model is not able to create the required "level playing field", and thus, that the quality label that predictions receive through a generic classification based on values for $\mathrm{MSESS}_{\mathrm{ini}}$ lacks general validity.

The mean-squared error MSE_r of the zero change reference prediction is given by the mean-squared difference between the observed bathymetries at the time of skill evaluation and start of the simulation. Thus, $\mathrm{MSESS}_{\mathrm{ini}}$ depends on the normalization of the error in the bed levels by the observed cumulative change away from the initial bed (Eq. 1.2). When MSE_r is progessively increasing in time, as can be expected for a trend, a constant skill throughout the simulation is obtained if the MSE of the predictions is a constant fraction of the MSE_r. In this way, the cumulative nature of morphology in combination with the zero change model provides a built-in, progressive lowering of the (metaphorical) bar. This is very different for predictands such as wave heights or precipitation, which are instantaneous values rather than accumulated quantities over the entire simulation duration. One may argue that the progressive relaxation of the stringency of the test qualitatively agrees with a modeller's intuition that the difficulty of the prediction situation increases with the prediction horizon. It is debatable, however, whether for a long-yearly trend, the prediction that nothing will change sets an ambitious enough quality standard.

Also, this interpretation, is not consistent with the fact that for seasonal systems, the zero change reference model does not exhibit a similar lowering of the (metaphorical) bar over the course of multiple years, since now the variation of MSE_r is bounded, regardless of the amount of gross change. As a consequence, the simulation of the progressive development has an unfair advantage over the simulation of the seasonal system and increasingly so further in time. Likewise, of two systems with the same net change, the system with larger gross changes—due to for instance an episodic event and subsequent partial recovery—can be expected to be the more difficult prediction situation, even though cumulative (net) changes from the reference cannot discern between the two situations. Clearly, the net change from the initial bed lacks information on the nature of the morphological development prior to the evaluation (trend, cyclic, episodic).

As a further consequence, the temporal evolution of model skill for seasonal systems may demonstrate an undesirable seasonal cycle; when a model, aimed at mimicking multiple summer–winter profile cycles, is initialized from a winter profile, a higher accuracy is required to obtain a certain skill level in the winter seasons, when the net observed changes and thus MSE_r are smaller, than in the summer seasons for which MSE_r is larger. This may explain the low skill scores in van Rijn et al. (2003) for the seasonal morphology at Egmond for periods in which the breaker bar is relatively stable.

For long-yearly simulations with process-based morphodynamic area models, there have been reports of skill scores steadily increasing with time (Dam et al., 2013, 2015, 2016; Luijendijk et al., 2017). A similar time-variation of skill was noticed in our analysis of a 15 year hindcast of the evolution of the Bornrif (Bosboom et al., 2014, i.e. Ch. 2). Upon comparison of the Bornrif's yearly bed changes with the cumulative bed changes from the start of the simulation, we noticed that the simulations capture little of the year-to-year variability while the larger-scale fields of cumulative change are reasonably well predicted. This was confirmed by an alternative skill computation that resulted in significantly lower skill values by considering bed changes in a one-year period rather than cumulative change over multiple years. Based on these findings, it was hypothesized that the relatively low values of $\mathrm{MSESS}_{\mathrm{ini}}$ at the beginning of the Bornrif simulation are mainly due to unskilful smaller scales. Over time, the relative contribution of these smaller scales to the cumulative change, and thus to $\mathrm{MSESS}_{\mathrm{ini}}$, decreases and, consequently, the contribution of the more skilful, persistent larger-scale trend increases.

To exemplify this effect, we designed a simple example (Bosboom and Reniers, 2018, i.e. Ch. 3), which assumes the observed and predicted anomalies to consist of two spatial scales of cumulative change: a larger, persistent and a smaller, intermittent scale. The skill of both the smaller and longer scales alone is constant with time. Nonetheless, the skill of the combined signal increases from low scores at the beginning, dictated by the unskilful small scales, to higher scores towards the end, dictated by the more skilful longer scales. In conclusion, an increase in skill for longer prediction horizons may well be indicative of the emerging of the more skilful larger scales, without the skill on these scales necessarily increasing in time.

In relation with the above, we found that $\mathrm{MSESS}_{\mathrm{ini}}$ exhibits a lack of symmetry in the case of sediment budget errors. In Ch. 3, a simple case was considered with a nonzero sediment budget error γ', due to a change of the mean of the prediction that is not present in the observations. Upon reversal of o' and p', the sediment budget error also counts towards MSE_r, leading to a a higher skill score. Apparently, there is a reward (in terms of a higher skill score) if the observations show a mean trend.

The above findings not only challenge the comparability of $\mathrm{MSESS_{ini}}$ at different times in a simulation, but between different simulations as well. The skill values merely indicate the fraction of improvement of model results compared to a model that assumes the initial bed of a particular simulation to persist, but have only limited meaning in a comparative analysis.

7.2 Spatial validation methods

The undesirable properties of traditional point-wise metrics when applied to high-resolution predictions (Sects. 1.1 and 1.2) led us to formulate the second main research objective (Objective 2): to develop validation methods and corresponding performance metrics that take the spatial structure of morphological patterns into account. It was elaborated in six research questions and objectives (2.1 to 2.6, see Sect. 1.3), which were addressed in Ch. 4 (Bosboom and Reniers, 2014b), Ch. 5 (Bosboom et al., 2019) and Ch. 6 (Bosboom and Reniers, 2014a). The main findings pertaining to each of these six research questions and objectives are summarized in the next sections (Sects. 7.2.1 to 7.2.6).

7.2.1 Development of a field deformation method

Objective 2.1 was formulated as follows: develop a field deformation method suited for the validation of morphological patterns and formulate (an) appropriate error metric(s) to be used in conjunction with this method. This objective was addressed in Ch. 4 (Bosboom and Reniers, 2014b), which presents a diagnostic tool for morphodynamic model validation that explicitly takes the (dis)agreement in spatial patterns into account. It classifies as a field deformation method, since it deforms the predictions to fit the observations as well as possible. Our method employs an efficient, nonrigid (i.e. allowing for free-form deformations) image warping technique—in an implementation by Kroon and Slump (2009)—to find the smooth displacement field between predictions and observations that minimizes the point-wise squared error. The technique, named Demon's registration (Thirion, 1998), bears similarities to optical flow and can be considered as similar to a minimization of the sum of square image intensities between the deformed predictions and observations (Pennec et al., 1999). The result of the image matching or warping is a vector field of displacements, which can be regarded as a displacement error field. The difference between the deformed predictions and the observations can be considered as an intensity or amplitude error field.

Based on the displacement error field and the pre- and post-warp intensity error fields, two new metrics are developed: (1) a mean location error $\overline{D}$ that is distilled from the displacement vector field; and (2) a combined error metric $\mathrm{RMSE_w}$ that takes both location and intensity errors into account. The location error $\overline{D}$ weights the local (backward) displacement magnitudes with their effect on the reduction of the local squared error and can be considered as a (weighted) mean distance between the predicted and observed morphological fields. For a full appreciation of the quality of a prediction, it should be considered in concert with both the original point-wise error RMSE_0 and the point-wise error of the deformed predictions, RMSE_1, which measures the agreement between predictions and observations if a zero penalty applies for misplacements of features. In order to quantify the overall relative performance between predictions, a (subjective) weighting of these three metrics must be carried out.

Alternatively, the weighting is already provided by RMSE_w that combines all relevant information on location errors and pre- and post-warp intensity errors. The weighting procedure locally relaxes the requirement of an exact match to an extent determined by the local displacement magnitude. To this end, a user-defined, physically intuitive parameter D_max is introduced, which is dependent on the prediction situation and the goal of the simulation. It can be seen as the maximum distance over which morphological features may be displaced for the prediction to still get (some) credit for predicting these features. Since it only requires a single, physically intuitive parameter, RMSE_w provides a robust basis for comparison.

The image matching optimizes the location of pixels with given predicted intensities (i.e. depth values) in an image and is therefore probably closest to the visual validation by morphologists. By implication, the method is not sediment-conserving, as opposed to the optimization method of Ch. 5 (Bosboom et al., 2019), which moves sediment rather than depth values.

7.2.2 Behaviour of displacement-based error metrics

Question 2.2 was formulated as follows: what is the behaviour of the error metric(s) as referred to in Objective 2.1, in comparison to the behaviour of point-wise metrics. This research question deals with the behaviour of the two metrics introduced in Ch. 4 (Bosboom and Reniers, 2014b), viz. the location error $\overline{D}$ and the RMSE_w, which takes both location and intensity errors into account. (Sect. 7.2.1).

Based on theoretical considerations, RMSE_w was seen to credit predictions to the degree that a larger error reduction can be obtained with smaller displacements. By definition, $\mathrm{RMSE}_1 \leqslant \mathrm{RMSE}_\mathrm{w} \leqslant \mathrm{RMSE}_0$. In fact, RMSE_w reduces to the pre-warp RMSE_0 if all displacements are larger than a user-defined D_max and to the post-warp RMSE_1 for displacements that are negligible relative to D_max. This aligns with the tendency of coastal morphologists to credit a prediction for the reproduction of features, albeit displaced, while imposing a relatively small penalty for misplacement. The intuitive weighting of these two aspects is mimicked by the user-defined parameter D_max.

To further answer Question 2.2, the new error metrics were used to diagnose the correspondence between model-generated pairs of morphological patterns as well as the relative ranking between the pairs. The fields were generated for the idealized case of a tidal inlet developing from an initially highly schematized geometry (Roelvink, 2006). First, we demonstrated, using a subset of the model-generated depth fields, which only differ with respect to the latitude and, hence, Coriolis parameter, that the location error $\overline{D}$ is able to capture the overall misplacement of the morphological patterns. Next, for a different series of depth-fields, the combined error metric $\mathrm{RMSE_w}$ was shown to outperform the conventional validation approach based on a strictly point-wise metric such as $\mathrm{RMSE_0}$, by avoiding the double penalty effect for misplaced features. The values for $\mathrm{RMSE_0}$, $\overline{D}$ and $\mathrm{RMSE_1}$ served to explain and support the ranking based on $\mathrm{RMSE_w}$. It was shown that, as opposed to the traditional $\mathrm{RMSE_0}$, $\mathrm{RMSE_w}$ makes choices as to which of two predictions is better, which are consistent with visual validation by experts.

7.2.3 An optimal transport method for morphological fields

Objective 2.3 was formulated as follows: develop an optimal transport method for the validation of morphological patterns and derive (a) corresponding error metric(s). This objective was addressed in Ch. 5 (Bosboom et al., 2019), which presents a diagnostic tool for morphodynamic model validation that defines the mismatch between predictions and observations in terms of a corrective sediment transport field. This optimal sediment transport field moves the misplaced sediment from the predicted to the observed morphology at the “cheapest” quadratic transportation cost. It is relatively easily found by solving an elliptic partial differential equation, viz. a Poisson equation, for which we have used the functions from the Matlab Partial Differential Equation (PDE) Toolbox.

A new domain-averaged error metric, the root-mean-squared transport error (RMSTE), is defined as the root-mean-square of the optimal transport field. By penalizing the sediment transport required for a match with the observations, the spatial structure of the error is taken into account; the RMSTE is sensitive to the volumes of misplaced sediment as well as to the distance over which this sediment must be transported.

The choice of a quadratic cost function was a pragmatic one: the exponent $p = 2$ in the cost function allows a relatively easy solution, as opposed to $p = 1$. With $p = 2$, the transport magnitudes at different locations are weighted quadratically, so extremes are heavily penalized. This leads to somewhat smeared out transport patterns with curved transport pathways, as shown in Sects. 5.3 and 5.4.

Our method, which we named effective transport difference (ETD), is a variation to a partial differential equation approach to the Monge–Kantorovich L^2 optimal transport problem. As such, it employs an irrotationality condition for the optimal transport in order to reformulate a transport optimization problem in terms of a partial differential equation. However, whereas the L^2 Monge–Kantorovich problem penalizes the quadratic distance the transformation moves each bit of material, weighted by the material's mass, our quadratic cost function penalizes the squared sediment transport, i.e. mass times distance, herewith retaining the original physical Monge's interpretation in terms of work (see Rachev and Rüschendorf, 1998), albeit in a quadratic sense. New aspects are further that our model boundaries are, in principle, open to sediment, which allows a bias to exist between the two bathymetric fields. Only in the special case of a boundary that is physically closed for sediment, e.g. land boundaries, one may assume that the transport error on the boundary is known and zero and replace one or more of the free boundaries by a closed boundary.

The ETD method, which moves sediment rather than features, results in a perfect transformation of the predicted to the observed morphological field, whereas, due to bed level differences between corresponding features in the two fields, the image warp, which, roughly speaking, moves features, does not allow an exact match. In line with this, the RMSTE represents the minimum (squared) sediment transport or work required to bridge the deviations between the morphological patterns, whereas the RMSE$_\mathrm{w}$ (Sects. 7.2.1 and 7.2.2) captures the visual disagreement between morphological patterns. Advantages of the ETD method over the image warp are that the ETD is mass-conserving, parameter-free and symmetric, the optimal transport from observations to predictions being the inverse of the optimal transport from predictions to observations.

7.2.4 The behaviour of the RMSTE

Question 2.4 was formulated as follows: what is the behaviour of the error metric(s) as referred to in Objective 2.3, in comparison to the behaviour of pointwise metrics. In order to answer this question, the behaviour of the newly introduced RMSTE and root-mean-squared error (RMSE) were compared, in Ch. 5 (Bosboom et al., 2019), for simple 1D and 2D cases as well as for more realistic model-generated morphological fields for a tidal inlet. The latter included the sets of "observed" and predicted fields that the image warp was tested against (see Sect. 7.2.2). The RMSTE was found to be capable of discriminating among model results, which is an important requirement of any error metric, and to lead to a different judgement than the RMSE as to which of two predictions is better. Whereas the RMSE only measures the amount of misplaced sediment, and, hence, already penalizes small misplacements of features heavily, the RMSTE also takes the distance over which this sediment is misplaced into account. Hence, larger spatial scales in the bathymetric error fields, requiring larger corrective transport distances, are penalized heavier than shorter scales. This makes the RMSTE more suited to demonstrate the quality of a high-variability prediction than the RMSE.

For the simple 1D and 2D cases, this was reflected in the RMSTE increasing with the misplacement distance of correctly sized features—until, in the extreme, sediment exchanged across the model boundaries may lead to a lower RMSTE.

Also, for the simple cases, the RMSTE was seen to avoid the consistent favouring of flat bed predictions that the RMSE suffers from; whether or not the RMSTE is larger for a flat bed prediction than for a correctly sized but misplaced feature depends strongly on the situation. In line with this, it was shown that the RMSE can be minimized by severely underpredicting feature amplitudes of misplaced features, whereas the RMSTE is only mildly sensitive to this undesirable effect.

Similarly, for the tidal inlet, it was demonstrated that more localized sediment misplacements, due to, for instance, incorrect Coriolis deflections, are diagnosed with better RMSTE scores than misplacements similar in volume but over larger distances. Predictions that require large corrective transports over large distances obtain the largest RMSTE. For the cases considered, the $\mathrm{RMSE_w}$, based on the image warp, and the RMSTE led to the same ranking amongst the predictions, albeit for certain choices of the warp's user-defined parameter that limits the distance over which features may be displaced for the prediction to still get (some) credit for predicting these features. In contrast, the RMSTE does not allow such a parameter. Further, it was concluded that inspection of the corrective transport fields, underlying the RMSTE, may provide some guidance as to how the model should be improved.

7.2.5 Validation statistics at multiple scales

Objective 2.5 was formulated as follows: develop a scale-selective validation framework that resolves the spatial distribution of appropriate validation statistics for multiple scales. This objective was addressed in Ch. 6 (Bosboom and Reniers, 2014a), which presents a validation method for 2D morphological predictions that allows any metric to selectively address multiple spatial scales. Herewith, information is provided on the variation of model performance with spatial scale and within the model domain. This information is not provided by the traditional approach to morphodynamic model validation to compute a single-number validation metric, such as the MSE or an MSE-based skill score, for the entire 2D model domain or a limited number of subdomains. Also, neighbourhood or smoothing methods (Sect. 1.2), which compute summary statistics for progressively smoother fields as obtained by smoothing filters, do not provide information on the spatial variation of performance in the model domain.

Our scale-selective validation approach employs a smoothing filter in such a way that, in addition to the domain-averaged statistics, localized validation statistics and areal maps of prediction quality are obtained per scale. It bears similarities to localized data analysis (Fotheringham et al., 2002) in that it computes local validation metrics in a sliding window of progressively larger size, moving across the domain. Hence, the term “scale” as considered by this method refers to geographic extent or areal size of focus. Compared to possible alternative strategies to scale-selective model validation, e.g. using band-pass spatial filters, the method is easy to implement and apply, and the results are relatively easy to interpret. This makes it a tool that can be readily used for practical purposes.

The applied validation metrics—metrics pertaining to the structure and variance of the morphological pattern—are close to the intuitive judgement of morphologists. The correlation ρ_{po} between the predicted and observed bed levels is used as a normalized measure of the structural similarity. Further, we defined a normalized measure of the amplitude similarity η as a function of the ratio of the standard deviations of predictions and observations $\hat{\sigma} = \sigma_p/\sigma_o$. The parameter η is bounded and invariant under the exchange of predictions and observations, as opposed to $\hat{\sigma}$. Hence, over-prediction and under-prediction are equally penalized. The structural and amplitude similarity are combined in a pattern skill score S, in line with the skill score proposed by Taylor (2001).

The various statistics are calculated for a range of window sizes, leading to maps of amplitude similarity, structural similarity and pattern skill per scale. Other validation metrics can be used as well; the method allows *any* metric to selectively address multiple spatial scales. Also, the method could be supplemented with a bias term at the largest scale. It is important to note that the employed metrics are point-wise metrics, but do not suffer from the double penalty effect; neither η nor S rewards the underestimation of variability, as opposed to point-wise accuracy metrics, such as the MSE, and derived skill metrics, such as the MSESS.

7.2.6 Information provided by the scale-selective approach

Question 2.6 was formulated as follows: what information is provided by the scale-selective framework as mentioned in Objective 2.5 and what is the added value of addressing multiple scales? This question was answered in Ch. 6 (Bosboom and Reniers, 2014a) through application to measured and computed bathymetric fields for the Bornrif—a dynamic attached bar at the North-western edge of the Wadden Sea island of Ameland.

Strong spatial differences in structural and amplitude similarity and pattern skill were found. Further, due to amongst others small-scale observed features that are not (well) reproduced in the predictions, a lower domain-averaged prediction quality was found at the smaller scales than at the larger scales. In relation to this, it was found that smoothing out the high-resolution detail increases the skill of the results especially at the smaller scales, even though the smoothed bathymetries are less realistic looking than the full-resolution bathymetries.

Clearly, the scale-selective validation can be used to determine local differences in structural and amplitude similarity and pattern skill. Further, it can be used to determine the smallest scales with sufficient skill and establish the resolution at which model-data comparisons are ideally presented. Finally, it may provide direction for model development targeted specifically at certain morphological scales.

7.3 Recommendations

From the findings of this thesis, several recommendations for model validation can be derived. First, Sect. 7.3.1 presents recommendations concerning the establishment of a set of performance measures. Next, in Sect. 7.3.2, recommendations for further development of spatial validation metrics are discussed.

7.3.1 Towards a morphological model validation suite

Any error metric condenses a large amount of data into a single number, therewith highlighting certain aspects of model performance only. By implication, multiple metrics are generally required to provide an adequate picture of the quality of morphological model results. Therefore, we recommend that a combination of metrics is used in the validation of morphological models and that the weighting is determined by the goal of the simulation. Such a procedure requires a good understanding of the behaviour of the available error metrics.

The development of an established set of performance measures, in combination with a set of internationally agreed validation cases, would be an important step in raising the level of morphodynamic model validation. This demands an investment from the modelling community, starting with giving due attention to model validation in morphological model studies.

A set of error metrics may include point-wise accuracy metrics such as the MSE, RMSE or mean absolute error (MAE) and derived skill metrics. It must be realised that in the presence of often inevitable location (and timing) errors, these accuracy and skill metrics tend to penalize rather than reward the correct prediction of the variance of bed levels and bed changes. For the MSE and MSE-based skill score (MSESS), this is easily monitored through the Murphy–Epstein decomposition. Therefore, it is advocated that for a good interpretation of model performance, the MSE and MSESS are supplemented with the anomaly error components $\rho_{p'o'}$, $\sigma_{p'}/\sigma_{o'}$, α', β', γ' and ϵ' (Sects. 7.1.1 and 7.1.2). Also, a decomposition of the MSE in terms of bed levels, reporting at the minimum σ_p/σ_o, ρ_{po} and MSE_{bias} (Ch. 2), is helpful.

If the MSESS is used and a correction for measurement error is called for, we advise the use of the skill formulation according to Sutherland et al. (2004) rather than van Rijn et al. (2003), in combination with a skill classification scheme that is not adjusted for measurement error (Sect. 7.1.3).

Unfortunately, the commonly used $\text{MSESS}_{\text{ini}}$ a.k.a. the BSS, which measures the accuracy relative to a prediction of zero morphological change, cannot be used reliably to rank morphodynamic model performance (Sect. 7.1.4). At best, skill levels can be judged on a case-by-case basis, but even then the comparison of skill levels at different times in a simulation has limited meaning. It is therefore discouraged to rely heavily on the $\text{MSESS}_{\text{ini}}$ for the determination of morphological model quality. If the $\text{MSESS}_{\text{ini}}$ must be used, it is recommended that the temporal variation of not only the $\text{MSESS}_{\text{ini}}$ is reported and analyzed, but also of the MSE, the MSE_{ini} and the above mentioned error terms, following from the Murphy–Epstein decomposition.

In order to reflect the intuitive judgement of morphologists, a set of performance metrics must include a metric that accounts for the spatial interdependency of the observed and predicted fields (Sect. 7.2). The location error $\overline{D}$ and combined error metric RMSE_{w}, as determined with a field deformation method, are suitable to measure the visual closeness between morphological patterns (Sects. 7.2.1 and 7.2.2). Alternatively, the RMSTE can be used to project the error in terms of an optimal transport from predictions to observations (Sects. 7.2.3 and 7.2.4). It serves as a recommendation that the optimal transport method, as opposed to the field deformation method, is mass-conserving, parameter-free and symmetric.

Further, the RMSTE may be helpful in calibrating morphodynamic models with respect to the morphodynamic timescale. In a first calibration step, an automated calibration routine, which minimizes the RMSTE, may be able to determine the optimal global model settings, such as certain transport parameters, that merely affect the morphodynamic timescale. In a next step, a more detailed calibration of other parameters can be undertaken using multiple error metrics, amongst others the RMSE and the RMSTE.

In addition to the computation of error metrics for the entire 2D domain or a limited number of subdomains, the scale-selective approach can be used to provide information on the variation of model performance with spatial scale and within the model domain (Sects. 7.2.5 and 7.2.6). To this end, normalized measures for structural similarity ρ_{po} and amplitude similarity $\eta = f(\sigma_p/\sigma_o)$, can be used and combined into a pattern skill score S (Ch. 6).

As mentioned, however attractive the concept of skill or relative accuracy, a comparative analysis based on $\text{MSESS}_{\text{ini}}$, lacks general validity. Alternatives are not self-evident. For longer-range simulations of seasonal systems, a more appropriate naive prediction may be the initial or last observed state for the same season. Also for a trend, an alternative skill score can be formulated by considering bed changes in a certain time period (e.g. one year, see Sect. 2.3.5) rather than cumulative change over multiple years. Even though these alternative persistence models may improve the comparability of skill values at different times in a simulation, they do not provide the fair reference required for a comparison of predictions across different prediction situations. In the presence of a trend, a more appropriate naive model could be some estimate of the trend (e.g. Davidson et al., 2010, for coastline modelling), which can be expected to provide a more stringent test than the reference prediction that nothing will change. For 2D morphology, a similar approach would be far from trivial. Finally, on a case-by-case basis, a quite useful choice of reference is a benchmark prediction with a different model or different model settings (Lesser, 2009; Gerritsen et al., 2011; Hallin et al., 2019).

7.3.2 Further development of spatial validation metrics

In future studies, the behaviour of the RMSTE in a range of practical applications will need to be considered. In order to do so, a more robust implementation of the ETD is required in order to deal with arbitrary model domains. Further, we anticipate that valuable additional information can be extracted from the optimal transport fields by isolating the various scales in the transport fields, for instance using our scale-selective validation method of Ch. 6 (Bosboom and Reniers, 2014a).

The choice of $p = 2$ in the ETD optimization problem, leading to quadratic transport costs, has enabled a relatively straightforward solution procedure resulting in a rotation-free optimal transport. For $p = 1$ and a domain boundary closed to sediment, our formulation and the L^p Monge–Kantorovich problem are equivalent and correspond to the original Monge mass transfer, which guarantees the shortest possible weighted transport distance and smallest transport magnitude. Numerical methods for solving the L^1 problem exist (Benamou and Carlier, 2015), but are considerably more complex than our L^2 solution procedure. Nonetheless, it may be worthwile to explore possibilities to solve the L^1 optimization problem. Such an approach would lead to the introduction of a new error metric, the mean absolute transport error (MATE). The MATE is to the RMSTE as the MAE is to RMSE, with that difference that MATE is based on $\mathbf{q}_{L1}$ rather than $\mathbf{q}_{L2}$.

Bibliography

El kadi Abderrezzak, K., Paquier, A., 2009. One-dimensional numerical modeling of sediment transport and bed deformation in open channels. Water Resources Research 45, W05404. doi:10.1029/2008WR007134.

Achete, F.M., Luijendijk, A., Tonnon, P.K., Stive, M.J.F., de Schipper, M.A., 2011. Morphodynamics of the Ameland Bornrif: an analogue for the Sand Engine. MSc thesis TU Delft. URL: http://resolver.tudelft.nl/uuid:76aabdcf-c3da-4a45-9720-39d2702e5c29.

Anthes, R.A., 1983. Regional models of the atmosphere in middle latitudes. Monthly Weather Review 111, 1306–1335. doi:10.1175/1520-0493(1983)111<1306:RMOTAI>2.0.CO;2.

Arpe, K., Hollingsworth, A., Tracton, M.S., Lorenc, A.C., Uppala, S., Kållberg, P., 1985. The response of numerical weather prediction systems to FGGE level IIb data. part II: Forecast verifications and implications for predictability. Quarterly Journal of the Royal Meteorological Society 111, 67–101. doi:10.1002/qj.49711146703.

Baart, F., van Ormondt, M., van Thiel de Vries, J.S.M., van Koningsveld, M., 2016. Morphological impact of a storm can be predicted three days ahead. Computers and Geosciences 90, 17–23. doi:10.1016/j.cageo.2015.11.011.

Benamou, J.D., Brenier, Y., 2000. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. Numerische Mathematik 84, 375–393. doi:10.1007/s002110050002.

Benamou, J.D., Brenier, Y., Guittet, K., 2002. The Monge–Kantorovitch mass transfer and its computational fluid mechanics formulation. International Journal for Numerical methods in fluids 40, 21–30. doi:10.1002/fld.264.

Benamou, J.D., Carlier, G., 2015. Augmented Lagrangian methods for transport optimization, mean field games and degenerate elliptic equations. Journal of Optimization Theory and Applications 167, 1–26. doi:f7q4h5.

Bogachev, V.I., Kolesnikov, A.V., 2012. The Monge–Kantorovich problem: achievements, connections, and perspectives. Russian Mathematical Surveys 67, 785–890. doi:10.1070/RM2012v067n05ABEH004808.

Bosboom, J., Mol, M., Reniers, A.J.H.M., Stive, M.J.F., de Valk, C.F., 2019. Optimal sediment transport for morphodynamic model validation. Manuscript submitted for publication.

Bosboom, J., Reniers, A., 2014a. Scale-selective validation of morphodynamic models, in: Proceedings 34th International Conference on Coastal Engineering, Seoul, South-Korea, pp. 1911–1920. doi:10.9753/icce.v34.sediment.75.

Bosboom, J., Reniers, A., 2018. The deceptive simplicity of the Brier skill score, in: Kim, Y.C. (Ed.), Handbook of Coastal and Ocean Engineering, pp. 1639–1663. doi:10/c5tr.

Bosboom, J., Reniers, A.J.H.M., 2014b. Displacement-based error metrics for morphodynamic models. Advances in Geosciences 39, 37–43. doi:10.5194/adgeo-39-37-2014.

Bosboom, J., Reniers, A.J.H.M., Luijendijk, A.P., 2014. On the perception of morphodynamic model skill. Coastal Engineering 94, 112–125. doi:10.1016/j.coastaleng.2014.08.008.

Bougeault, P., 2003. The WGNE survey of verification methods for numerical prediction of weather elements and severe weather events. CAS/JSC WGNE Report, 18, WMO/TD-NO. 1173 Appendix C, 1–11. URL: http://www.wcrp-climate.org/documents/wgne18rpt.pdf.

Brenier, Y., 2003. Extended Monge–Kantorovich theory, in: Optimal transportation and applications. Springer Berlin Heidelberg, pp. 91–121. doi:10/c5vw.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Monthly Weather Review 78, 1–3. doi:10/fp62r6.

Brier, G.W., Allen, R.A., 1951. Verification of weather forecasts. Compendium of Meteorology , 841–848. doi:10/c5vx.

Briere, C., Giardino, A., van der Werf, J., 2011. Morphological modeling of bar dynamics with Delft3D: The quest for optimal free parameter settings using an automatic calibration technique. Coastal Engineering Proceedings 1, 60. doi:10.9753/icce.v32.sediment.60.

Casati, B., Wilson, L.J., Stephenson, D.B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E.E., Brown, B.G., Mason, S., 2008. Forecast verification: current status and future directions. Meteorological applications 15, 3–18. doi:10.1002/met.52.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development 7, 1247–1250. doi:10.5194/gmd-7-1247-2014.

Dam, G., van der Wegen, M., Labeur, R.J., Roelvink, D., 2016. Modeling centuries of estuarine morphodynamics in the Western Scheldt estuary. Geophysical Research Letters 43, 3839–3847. doi:10.1002/2015GL066725.

Dam, G., van der Wegen, M., Roelvink, D., 2013. Long-term performance of process-based models in estuaries, in: Proceedings Coastal Dynamics, Bordeaux, France, pp. 409–420.

Dam, G., van der Wegen, M., Roelvink, D.J.A., Labeur, R., Bliek, B., 2015. Simulation of long-term morphodynamics of the Western Scheldt, in: Proceedings IAHR World Congress, The Hague, The Netherlands.

Davidson, M.A., Lewis, R.P., Turner, I.L., 2010. Forecasting seasonal to multi-year shoreline change. Coastal Engineering 57, 620–629. doi:10.1016/j.coastaleng.2010.02.001.

Davidson, M.A., Splinter, K.D., Turner, I.L., 2013. A simple equilibrium model for predicting shoreline change. Coastal Engineering 73, 191–202. doi:10.1016/j.coastaleng.2012.11.002.

Dodet, G., Castelle, B., Masselink, G., Scott, T., Davidson, M., Floc'h, F., Jackson, D., Suanez, S., 2019. Beach recovery from extreme storm activity during the 2013–14 winter along the Atlantic coast of Europe. Earth Surface Processes and Landforms 44, 393–401. doi:10.1002/esp.4500.

Elmilady, H., van der Wegen, M., Roelvink, D., Jaffe, B.E., 2019. Intertidal area disappears under sea level rise: 250 years of morphodynamic modeling in San Pablo bay, California. Journal of Geophysical Research: Earth Surface 124, 38–59. doi:10.1029/2018JF004857.

Evans, L.C., 1997. Partial differential equations and Monge–Kantorovich mass transfer, in: Current Developments in Mathematics, International Press of Boston. pp. 65–126. URL: https://math.berkeley.edu/~evans/Monge-Kantorovich.survey.pdf.

Fortunato, A.B., Nahon, A., Dodet, G., Pires, A.R., Freitas, M.C., Bruneau, N., Azevedo, A., Bertin, X., Benevides, P., Andrade, C., Oliveira, A., 2014. Morphological evolution of an ephemeral tidal inlet from opening to closure: The Albufeira inlet, Portugal. Continental Shelf Research 73, 49–63. doi:10.1016/j.csr.2013.11.005.

Fotheringham, A.S., Brundson, C., Charlton, M., 2002. Geographically weighted regression: the analysis of spatially varying relationships. Wiley, Chichester, West Sussex, England.

Gallagher, E.L., Elgar, S., Guza, R.T., 1998. Observations of sand bar evolution on a natural beach. Journal of Geophysical Research: Oceans (1978–2012) 103, 3203–3215. doi:10.1029/97JC02765.

Ganju, N.K., Jaffe, B.E., Schoellhamer, D.H., 2011. Discontinuous hindcast simulations of estuarine bathymetric change: A case study from Suisun Bay, California. Estuarine, Coastal and Shelf Science 93, 142–150. doi:10.1016/j.ecss.2011.04.004.

Gerritsen, H., Sutherland, J., Deigaard, R., Sumer, M., Fortes, C.J., Sierra, J.P., Schmidtke, U., 2011. Composite modelling of interactions between beaches and structures. Journal of Hydraulic Research 49, 2–14. doi:10.1080/00221686.2011.589134.

Gilbert, G.K., 1884. Finley's tornado predictions. American Meteorological Journal. A Monthly Review of Meteorology and Allied Branches of Study (1884–1896) 1, 166–172.

Gilleland, E., Ahijevych, D., Brown, B.G., Casati, B., Ebert, E.E., 2009. Intercomparison of spatial forecast verification methods. Weather and Forecasting 24, 1416–1430. doi:10.1175/2009WAF2222269.1.

Gilleland, E., Ahijevych, D.A., Brown, B.G., Ebert, E.E., 2010a. Verifying Forecasts Spatially. Bulletin of the American Meteorological Society 91, 1365–1373. doi:10.1175/2010BAMS2819.1.

Gilleland, E., Lindström, J., Lindgren, F., 2010b. Analyzing the image warp forecast verification method on precipitation fields from the icp. Weather and Forecasting 25, 1249–1262. doi:10.1175/2010WAF2222365.1.

Guerin, T., Bertin, X., Chaumillon, E., 2016. Wave control on the rhythmic development of a wide estuary mouth sandbank: A process-based modelling study. Marine Geology 380, 79–89. doi:10.1016/j.margeo.2016.06.013.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology 377, 80–91. doi:10.1016/j.jhydrol.2009.08.003.

Haben, S., Ward, J.A., Vukadinovic Greetham, D., Singleton, C., Grindrod, P., 2014. A new error measure for forecasts of household-level, high resolution electrical energy consumption. International Journal of Forecasting 30, 246–256. doi:10.1016/j.ijforecast.2013.08.002.

Hallermeier, R.J., 1980. A profile zonation for seasonal sand beaches from wave climate. Coastal Engineering 4, 253 – 277. doi:10.1016/0378-3839(80)90022-8.

Hallin, C., Huisman, B.J., Larson, M., Walstra, D.J.R., Hanson, H., 2019. Impact of sediment supply on decadal-scale dune evolution — Analysis and modelling of the Kennemer dunes in the Netherlands. Geomorphology 337, 94–110. doi:10.1016/j.geomorph.2019.04.003.

Henderson, S.M., Allen, J.S., Newberger, P.A., 2004. Nearshore sandbar migration predicted by an eddy-diffusive boundary layer model. Journal of Geophysical Research: Oceans 109, C06024. doi:10.1029/2003JC002137.

Hollingsworth, A., Arpe, K., Tiedtke, M., Capaldo, M., Savijärvi, H., 1980. The Performance of a Medium-Range Forecast Model in Winter – Impact of Physical Parameterizations . Monthly Weather Review 108, 1736–1773. doi:10.1175/1520-0493(1980)108<1736:TPOAMR>2.0.CO;2.

Jolliffe, I.T., Stephenson, D.B.T.A., 2012. Forecast verification: a practitioner's guide in atmospheric science. 2nd (304 pages) ed., John Wiley & Sons,. doi:10.1002/9781119960003.

Keil, C., Craig, G.C., 2009. A displacement and amplitude score employing an optical flow technique. Weather and Forecasting 24, 1297–1308. doi:10.1175/2009WAF2222247.1.

Koh, T.Y., Wang, S., Bhatt, B.C., 2012. A diagnostic suite to assess NWP performance. Journal of Geophysical Research 117, D13109. doi:10.1029/2011JD017103.

Kroon, D.J., Slump, C.H., 2009. MRI modality transformation in demon registration, in: From Nano to Macro., IEEE International Symposium on Biomedical Imaging. pp. 963–966. doi:10.1109/ISBI.2009.5193214.

Lesser, G.R., 2009. An approach to medium-term coastal morphological modelling. PhD thesis, IHE Delft Institute for Water Education. URL: http://resolver.tudelft.nl/uuid:27a1ffa0-580e-4eae-907b-ce6f901e652e.

Lesser, G.R., Roelvink, J.A., van Kester, J.A.T.M., Stelling, G., 2004. Development and validation of a three-dimensional morphological model. Coastal Engineering 51, 883–915. doi:10.1016/j.coastaleng.2004.07.014.

Livezey, R.E., Hoopingarner, J.D., Huang, J., 1995. Verification of official monthly mean 700-hPa height forecasts: An update. Weather and Forecasting 10, 512–527. doi:10.1175/1520-0434(1995)010<0512:VOOMMH>2.0.CO;2.

Luijendijk, A.P., Ranasinghe, R.W.M.R.J.B., Huisman, B., de Schipper, M.A., Swinkels, C., Walstra, D.J.R., Stive, M.J.F., 2017. The initial morphological response of the Sand Engine: a process-based modelling study. Coastal engineering 119, 1–14. doi:10.1016/j.coastaleng.2016.09.005.

Luijendijk, A.P., de Schipper, M.A., Ranasinghe, R., 2019. Morphodynamic acceleration techniques for multi-timescale predictions of complex sandy interventions. Journal of Marine Science and Engineering 7. doi:10.3390/jmse7030078.

Marzban, C., Sandgathe, S., 2010. Optical flow for verification. Weather and Forecasting 25, 1479–1494. doi:10.1175/2010WAF2222351.1.

Mass, C.F., Ovens, D., Westrick, K., Colle, B.A., 2002. Does increasing horizontal resolution produce more skillful forecasts? Bulletin of the American Meteorological Society 83, 407–430. doi:10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

Mathworks, 2015. Matlab and partial differential equation toolbox release 2015b. The MathWorks, Inc., Natick, Massachusetts, United States.

McCall, R.T., Masselink, G., Poate, T.G., Roelvink, J.A., Almeida, L.P., 2015. Modelling the morphodynamics of gravel beaches during storms with XBeach-G. Coastal Engineering 103, 52–66. doi:doi.org/10.1016/j.coastaleng.2015.06.002.

McCall, R.T., de Vries, J.S.M.V.T., Plant, N.G., van Dongeren, A.R., Roelvink, J.A., Thompson, D.M., Reniers, A.J.H.M., 2010. Two-dimensional time dependent hurricane overwash and erosion modeling at Santa Rosa island. Coastal Engineering 57, 668–683. doi:10.1016/j.coastaleng.2010.02.006.

Minneboo, F.A.J., 1995. Jaarlijkse Kustmetingen: Richtlijnen voor de inwinning, bewerking en opslag van gegevens van jaarlijkse kustmetingen. Technical Report. RIKZ-95.022, Ministry of Transport, Public Works and Water Management. URL: http://resolver.tudelft.nl/uuid:76f2634d-f3c4-4609-aa4c-44d641da28f1. (in Dutch).

Mol, M., Bosboom, J., De Valk, C.F., Reniers, A.J.H.M., Stive, M.J.F., Yuan, J., 2015. The Effective Transport Difference: a new concept for morphodynamic model validation. MSc thesis TU Delft. URL: http://resolver.tudelft.nl/uuid:aa35b28a-8bca-4b33-b99c-05a79e9154a6.

Monge-Ganuzas, M., Gainza, J., Liria, P., Epelde, I., Uriarte, A., Garnier, R., González, M., Nuñez, P., Jaramillo, C., Medina, R., 2017. Morphodynamic evolution of Laida beach (Oka estuary, Urdaibai Biosphere Reserve, southeastern Bay of Biscay) in response to supratidal beach nourishment actions. Journal of Sea Research 130, 85–95. doi:10.1016/j.seares.2017.06.003.

Mosselman, E., Le, T.B., 2016. Five common mistakes in fluvial morphodynamic modeling. Advances in Water Resources: Part A 93, 15–20. doi:10.1016/j.advwatres.2015.07.025.

Murphy, A.H., 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. Monthly Weather Review 116, 2417–2424. doi:10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.

Murphy, A.H., 1992. Climatology, persistence, and their linear combination as standards of reference in skill scores. Weather and Forecasting 7, 692–698. doi:10.1175/1520-0434(1992)007<0692:CPATLC>2.0.CO;2.

Murphy, A.H., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather and forecasting 8, 281–293. doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Murphy, A.H., 1996a. The Finley affair: A signal event in the history of forecast verification. Weather and Forecasting 11, 3–20. doi:10.1175/1520-0434(1996)011<0003:TFAASE>2.0.CO;2.

Murphy, A.H., 1996b. General decompositions of MSE-based skill scores: Measures of some basic aspects of forecast quality. Monthly Weather Review 124, 2353–2369. doi:10.1175/1520-0493(1996)124<2353:GDOMBS>2.0.CO;2.

Murphy, A.H., Epstein, E.S., 1989. Skill scores and correlation coefficients in model verification. Monthly Weather Review 117, 572–582. doi:10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2.

Novaczek, E., Devillers, R., Edinger, E., 2019. Generating higher resolution regional seafloor maps from crowd-sourced bathymetry. PLOS ONE 14, 1–23. doi:10.1371/journal.pone.0216792.

Oberkampf, W.L., Trucano, T.G., 2002. Verification and validation in computational fluid dynamics. Progress in Aerospace Sciences 38, 209–272. doi:10.1016/S0376-0421(02)00005-2.

Orzech, M.D., Reniers, A.J.H.M., Thornton, E.B., MacMahan, J.H., 2011. Megacusps on rip channel bathymetry: Observations and modeling. Coastal Engineering 58, 890–907. doi:10.1016/j.coastaleng.2011.05.001.

Pedrozo-Acuña, A., Simmonds, D.J., Otta, A.K., Chadwick, A.J., 2006. On the cross-shore profile change of gravel beaches. Coastal Engineering 53, 335–347. doi:10.1016/j.coastaleng.2005.10.019.

Pennec, X., Cachier, P., Ayache, N., 1999. Understanding the “demons algorithm”: 3d non-rigid registration by gradient descent, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI’99, Springer. pp. 597–605. doi:crh7nn.

Plant, N.G., Holland, K.T., Puleo, J.A., Gallagher, E.L., 2004. Prediction skill of nearshore profile evolution models. Journal of Geophysical Research: Oceans 109. doi:10.1029/2003JC001995.

Rachev, S.T., Rüschendorf, L., 1998. Mass Transportation Problems: Volume I: Theory. Springer Science & Business Media. doi:10.1007/b98893.

Radermacher, M., de Schipper, M.A., Reniers, A.J.H.M., 2018. Sensitivity of rip current forecasts to errors in remotely-sensed bathymetry. Coastal Engineering 135, 66–76. doi:10.1016/j.coastaleng.2018.01.007.

van Rijn, L.C., Walstra, D.J.R., Grasmeijer, B.T., Sutherland, J., Pan, S., Sierra, J.P., 2003. The predictability of cross-shore bed evolution of sandy beaches at the time scale of storms and seasons using process-based profile models. Coastal Engineering 47, 295–327. doi:10.1016/S0378-3839(02)00120-5.

Roberts, N.M., Lean, H.W., 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Monthly Weather Review 136, 78–97. doi:10.1175/2007MWR2123.1.

Roelvink, D., Reniers, A., 2012. A guide to Modeling Coastal Morphology. volume 12. World Scientific Publishing Company. doi:10.1142/7712.

Roelvink, D., Reniers, A., van Dongeren, A., van Thiel de Vries, J., McCall, R., Lescinski, J., 2009. Modelling storm impacts on beaches, dunes and barrier islands. Coastal Engineering 56, 1133–1152. doi:10.1016/j.coastaleng.2009.08.006.

Roelvink, J.A., 2006. Coastal morphodynamic evolution techniques. Coastal Engineering 53, 277–287. doi:10.1016/j.coastaleng.2005.10.015.

Ruessink, B.G., Kuriyama, Y., 2008. Numerical predictability experiments of cross-shore sandbar migration. Geophysical Research Letters 35, L01603. doi:10.1029/2007GL032530.

Ruessink, B.G., Kuriyama, Y., Reniers, A.J.H.M., Roelvink, J.A., Walstra, D.J.R., 2007. Modeling cross-shore sandbar behavior on the timescale of weeks. Journal of Geophysical Research: Earth Surface 112, 2003–2012. doi:10.1029/2006JF000730.

Ruggiero, P., Walstra, D.J.R., Gelfenbaum, G., van Ormondt, M., 2009. Seasonal-scale nearshore morphological evolution: Field observations and numerical modeling. Coastal Engineering 56, 1153–1172. doi:10.1016/j.coastaleng.2009.08.003.

Santambrogio, F., 2015. Optimal transport for applied mathematicians. Birkäuser, NY , 99–102. doi:10.1007/978-3-319-20828-2.

Scott, T.R., Mason, D.C., 2007. Data assimilation for a coastal area morphodynamic model: Morecambe bay. Coastal Engineering 54, 91–109. doi:10.1016/j.coastaleng.2006.08.008.

Simmons, J.A., Harley, M.D., Marshall, L.A., Turner, I.L., Splinter, K.D., Cox, R.J., 2017. Calibrating and assessing uncertainty in coastal numerical models. Coastal Engineering 125, 28–41. doi:10.1016/j.coastaleng.2017.04.005.

Simmons, J.A., Splinter, K.D., Harley, M.D., Turner, I.L., 2019. Calibration data requirements for modelling subaerial beach storm erosion. Coastal Engineering 152, 103507. doi:10.1016/j.coastaleng.2019.103507.

Stive, M.J.F., de Schipper, M.A., Luijendijk, A.P., Aarninkhof, S.G.J., van Gelder-Maas, C., van Thiel de Vries, J.S.M., de Vries, S., Henriquez, M., Marx, S., Ranasinghe, R., 2013. A new alternative to saving our beaches from sea-level rise: the sand engine. Journal of Coastal Research 29, 1001–1008. doi:10.2112/JCOASTRES-D-13-00070.1.

Sutherland, J., Peet, A.H., Soulsby, R.L., 2004. Evaluating the performance of morphological models. Coastal Engineering 51, 917–939. doi:10.1016/j.coastaleng.2004.07.015.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres 106, 7183–7192. doi:10.1029/2000JD900719.

Taylor, R., 1990. Interpretation of the correlation coefficient: A basic review. Journal of diagnostic medical sonography 6, 35–39. doi:10.1177/875647939000600106.

Thacker, B.H., Doebling, S.W., Hemez, F.M., Anderson, M.C., Pepin, J.E., Rodriguez, E.A., 2004b. Concepts of Model Verification and Validation. Technical Report, Los Alamos National Lab., Los Alamos, NM (US) doi:10.2172/835920.

Thirion, J.P., 1998. Image matching as a diffusion process: an analogy with Maxwell's demons. Medical Image Analysis 2, 243–260. doi:10.1016/S1361-8415(98)80022-4.

Villani, C., 2003. Topics in optimal transportation. 58, American Mathematical Soc. URL: https://bookstore.ams.org/gsm-58/.

Walstra, D.J.R., Reniers, A.J.H.M., Ranasinghe, R., Roelvink, J.A., Ruessink, B.G., 2012. On bar growth and decay during interannual net offshore migration. Coastal Engineering 60, 190–200. doi:10.1016/j.coastaleng.2011.10.002.

Wang, Z., Simoncelli, E.P., Bovik, A.C., 2003. Multiscale structural similarity for image quality assessment, in: Proceedings 37th Conference on Sigals, Systems and Computers, pp. 1398–1402. doi:10.1109/ACSSC.2003.1292216.

van der Wegen, M., Jaffe, B.E., Roelvink, J.A., 2011. Process-based, morphodynamic hindcast of decadal deposition patterns in San Pablo Bay, California, 1856–1887. Journal of Geophysical Research 116, F02008. doi:10.1029/2009JF001614.

van der Wegen, M., Roelvink, J., 2012. Reproduction of estuarine bathymetry by means of a process-based model: Western Scheldt case study, the Netherlands. Geomorphology 179, 152–167. doi:10.1016/j.geomorph.2012.08.007.

Wiegman, N., Perluka, R., Oude Elberink, S., Vogelzang, J., 2005. Vaklodingen: de inwintechnieken en hun combinaties. Vergelijking tussen verschillende inwintechnieken en de combinaties ervan. Technical Report. AGI-2005-GSMH-012. Adviesdienst Geo-Informatica en ICT (AGI). Delft. (in Dutch).

Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences. volume 100. 3rd ed., Academic Press.

Williams, C.N., Cornford, S.L., Jordan, T.M., Dowdeswell, J.A., Siegert, M.J., Clark, C.D., Swift, D.A., Sole, A., Fenty, I., Bamber, J.L., 2017. Generating synthetic fjord bathymetry for coastal Greenland. The Cryosphere 11, 363–380. doi:10.5194/tc-11-363-2017.

Williams, J.J., de Alegría-Arzaburu, A.R., McCall, R.T., van Dongeren, A., 2012. Modelling gravel barrier profile response to combined waves and tides using XBeach: Laboratory and field results. Coastal Engineering 63, 62–80. doi:10.1016/j.coastaleng.2011.12.010.

Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bulletin of the American Meteorological Society 63, 1309–1313. doi:10.1175/1520-0477(1982)063<1309:SCOTEO>2.0.CO;2.

Winkler, R.L., 1994. Evaluating probabilities: Asymmetric scoring rules. Management Science 40, 1395–1405. doi:10.1287/mnsc.40.11.1395.

Winkler, R.L., Muñoz, J., Cervera, J.L., Bernardo, J.M., Blattenberger, G., Kadane, J.B., Lindley, D.V., Murphy, A.H., Oliver, R.M., Ríos-Insua, D., 1996. Scoring rules and the evaluation of probabilities. Test 5, 1–60. doi:10.1007/BF02562681.

Ziegeler, S.B., Dykes, J.D., Shriver, J.F., 2012. Spatial error metrics for oceanographic model verification. Journal of Atmospheric and Oceanic Technology 29, 260–266. doi:10.1175/JTECH-D-11-00109.1.

Curriculum vitae

Personalia

Judith Bosboom, born 5 June 1970 in Nijmegen

Education

1982–1988	VWO, Petrus Canisius College, Alkmaar (pre-university education)
1989–1995	MSc Civil Engineering, Delft University of Technology (cum laude)
2006–2010	Photoacademy, Amsterdam

Awards

1996	Best MSc thesis of the Faculty of Civil Engineering
2000	Most Promising Student Award, International Association of Dredging Companies
2011–2019	Best lecturer of the MSc Hydraulic Engineering (7 awards)
2016	Best lecturer of Civil Engineering
2016	Best lecturer of Delft University of Technology

Employment

1991–1994	Delft University of Technology
1995–2000	Delft Hydraulics (now part of Deltares)
2000–2002	Turner, management consultancy
2002–2005	P2 managers, process and project management
2006–2007	City of Rotterdam
2006–2011	Judith Bosboom Photography and Text
2007–	Delft University of Technology

Publications

Refereed scientific papers and book chapters

Bosboom, J., Mol, M., Reniers, A.J.H.M., Stive, M.J.F., de Valk, C.F., 2019. Optimal sediment transport for morphodynamic model validation. Manuscript submitted for publication.

Bosboom, J., Reniers, A., 2018. The deceptive simplicity of the Brier skill score, in: Kim, Y.C. (Ed.), Handbook of Coastal and Ocean Engineering, pp. 1639–1663. doi:10/c5tr.

Bosboom, J., Reniers, A.J.H.M., Luijendijk, A.P., 2014. On the perception of morphodynamic model skill. Coastal Engineering 94, 112–125. doi:10.1016/j.coastaleng.2014.08.008.

Bosboom, J., Reniers, A.J.H.M., 2014. Displacement-based error metrics for morphodynamic models. Advances in Geosciences 39, 37–43. doi:10.5194/adgeo-39-37-2014.

Ranasinghe, R., Swinkels, C., Luijendijk, A., Roelvink, D., Bosboom, J., Stive, M., Walstra, D., 2011. Morphodynamic upscaling with the MORFAC approach: Dependencies and sensitivities. Coastal engineering 58, 806–811. doi:10.1016/j.coastaleng.2011.03.010.

de Meijer, R.J., Bosboom, J., Cloin, B., Katopodi, I., Kitou, N., Koomans, R.L., Manso, F., 2002. Gradation effects in sediment transport. Coastal Engineering 47, 179–210. doi:10.1016/S0378-3839(02)00125-4.

Bosboom, J., Klopman, G., Roelvink, J.A., Battjes, J.A., 1997. Boussinesq modelling of wave-induced horizontal particle velocities. Coastal engineering 32, 163–180. doi:10.1016/S0378-3839(97)81748-6.

Non-refereed scientific papers

Bosboom, J., Reniers, A., 2014. Scale-selective validation of morphodynamic models, in: Proceedings 34th International Conference on Coastal Engineering, Seoul, South-Korea, pp. 1911–1920. doi:10.9753/icce.v34.sediment.75.

Villani, M., Bosboom, J., Zijlema, M., Stive, M.J.F., 2012. Circulation patterns and shoreline response induced by submerged breakwaters, in: Proceedings 33rd International Conference on Coastal Engineering, Santander, Spain. doi:10.9753/icce.v33.structures.25.

Ranasinghe, R., Bosboom, J., Uhlenbrook, S., Roelvink, D., Ngo, H.Q., Stive, M., 2011. A scale aggregated model to estimate climate change driven coastline change along inlet interrupted coasts, in: Proceedings Coastal Sediments 2011, Miami, US, pp. 286–298. doi:bj9x68.

Ranasinghe, R., Swinkels, C., Luijendijk, A., Bosboom, J., Roelvink, D., Stive, M., Walstra, D., 2010. Morphodynamic upscaling with the MORFAC approach, in: Proceedings 32nd International Conference on Coastal Engineering, Shanghai, China. doi:10.9753/icce.v32.sediment.59.

Bosboom, J., Klopman, G., 2000. Intra-wave sediment transport modelling, in: Proceedings 27th International Conference on Coastal Engineering, Sydney, Australia, pp. 2453–2466. doi:10.1061/40549(276)192.

Bosboom, J., 2000. Wind-wave induced oscillatory velocities predicted by Boussinesq models. Terra et Aqua 80, 12–20.

Stive, M.J.F., Cloin, B., Jiménez, J., Bosboom, J., 1999. Long-term cross-shoreface sediment fluxes, in: Proceedings Coastal Sediments 1999, New York, US, ASCE. pp. 505–518.

Koomans, R.L., Bosboom, J., de Meijer, R.J., Venema, L.B., 1999. Effects of density on cross-shore sediment transport, in: Proceedings Coastal Sediments 1999, New York, US, ASCE. pp. 313–324.

Bosboom, J., Reniers, A., van Dongeren, A., 1999b. How numerical models can contribute to experimental research, in: Hydralab workshop on experimental research and synergy effects with mathematical models, Hannover, Germany, ASCE. pp. 181–188.

Bosboom, J., Koomans, R., Reniers, A., 1999a. Laboratory experiments on suspended sediment concentration and fluxes, in: Proceedings Coastal Sediments 1999, New York, US, ASCE. pp. 179–194.

Chatelus, Y., Katopodi, I., Dohmen-Janssen, M., Ribberink, J.S., Samothrakis, P., Cloin, B., Savioli, J.C., Bosboom, J., O'Connor, B.A., Hein, R., Hamm, L., 1998. Size gradation effects in sediment transport, in: Proceedings 26th International Conference on Coastal Engineering, Copenhagen, Denmark, pp. 2435–2448. doi:10.1061/9780784404119.183.

Bosboom, J., Klopman, G., Reniers, A., Stive, M.J.F., 1998. Analytical model for wave-related transport, in: Proceedings 26th International Conference on Coastal Engineering, Copenhagen, Denmark, pp. 2573–2586. doi:10.1061/9780784404119.194.

Bosboom, J., Klopman, G., Roelvink, J.A., Battjes, J.A., 1996. Wave kinematics computations using Boussinesq models, in: Proceedings 25th International Conference on Coastal Engineering, Orlando, USA., pp. 109–122. doi:10.1061/9780784402429.009.

Books

Bosboom, J., Stive, M.J.F., 2015. Coastal Dynamics I: lectures notes CIE4305. 5th (584 pages) ed., Delft Academic Press, 1st ed. 2010. URL: https://www.delftacademicpress.nl/f019.php.

Bosboom, J., Fuchs, M., 2006. Waterproef: Vier persoonlijke geschiedenissen van de waterbouw. Water Research Centre Delft. URL: http://resolver.tudelft.nl/uuid:caa80e35-1425-4505-8be6-ded1d7110201. (in Dutch).

THIS DOCTORAL THESIS IS ABOUT

The behaviour of the widely used mean-squared-error skill score with the initial bed as the reference, which goes by the name Brier skill score.

The development of novel validation methods and corresponding error metrics that take the spatial structure of morphological patterns into account:

1. *A field deformation or warping method*, which deforms the predictions as to minimize the misfit with observations;
2. *An optimal transport method*, which moves misplaced sediment from the predicted to the observed morphology through an optimal, rotation-free sediment transport field;
3. *A scale-selective validation approach*, which allows any metric to selectively address multiple spatial scales.

AND CONTAINS THE FINDINGS

The use of a single performance metric leads to an inadequate interpretation of quality.

A set of performance metrics for morphological models must include a metric—such as the root-mean-squared transport error (RMSTE)—that takes the spatial structure of morphological patterns into account.

Optimizing the mean-squared error (MSE) or derived skill score (MSESS or BSS) of a morphological prediction leads to undesired underprediction of the variance of bed changes.

The MSE-based skill score using the initial bed as the reference (a.k.a. the BSS) fails at making predictions comparable, whether across different prediction situations or across different times in a simulation.

ISBN 978-94-6384-091-0

An electronic version of this thesis is available at *https://repository.tudelft.nl*