# On the perception of morphodynamic model skill

J. Bosboom[a,*], A.J.H.M. Reniers[a,b], A.P. Luijendijk[c,a]

[a] *Faculty of Civil Engineering and Geosciences, Delft University of Technology, Department of Hydraulic Engineering, P.O. Box 5048, 2600 GA Delft, The Netherlands*
[b] *Deltares, Unit Marine and Coastal Systems, P.O. Box 177, 2600 MH Delft, The Netherlands*
[c] *Deltares, Unit Hydraulic Engineering, P.O. Box 177, 2600 MH Delft, The Netherlands*

## 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.

*Keywords:*
Brier skill score, Mean-squared error, Model skill, Morphodynamic modelling, Model validation, Zero change model

## 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 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 time-scales 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

[*] Corresponding author. Tel.: +31 15 27 84606; fax: +31 15 27 85124
*Email addresses:* `j.bosboom@tudelft.nl` (J. Bosboom), `a.j.h.m.reniers@tudelft.nl` (A.J.H.M. Reniers), `arjen.luijendijk@deltares.nl` (A.P. Luijendijk)

[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).

*Author manuscript accepted 18 August 2014 for publication in Coastal Engineering* doi: 10.1016/j.coastaleng.2014.08.008

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, 2014a). 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, section 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 section 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 section 4, the implications for morphological model validation are discussed. Finally, section 5 presents conclusions and discusses avenues for adaptation of validation strategies.

## 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, section 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 section 2.2, the metrics are interpreted in the context of the validation of morphological fields, using the initial bed as the reference. Third, section 2.3 discusses the impact of the zero change reference model on the perception of morphodynamic model skill. Finally, section 2.4 demonstrates that the MSESS tends to reward an underestimation of the variance of bed changes.

### *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):

$$\mathrm{ESS} = \frac{E - E_r}{E_i - E_r} \tag{1}$$

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. (1) yields (Murphy, 1988):

$$\mathrm{MSESS} = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE}_r} \tag{2}$$

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:

$$\mathrm{MSE} = \langle (p - o)^2 \rangle = \frac{1}{n} \sum_i^n w_i \, (p_i - o_i)^2 \tag{3}$$

where the angle brackets denote spatially weighted averaging, $(p_i, o_i)$ are the $i$th 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. (3) upon substitution as:

$$\mathrm{MSE} = \langle (p' - o')^2 \rangle. \tag{4}$$

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

$$\mathrm{MSE}_r = \langle (r - o)^2 \rangle = \langle o'^2 \rangle. \tag{5}$$

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 A). Herewith, Equation (4) can be written as (cf. eqs. (A.2) and (A.3)):

$$\mathrm{MSE} = \sigma_{o'}^2 (1 - \alpha' + \beta' + \gamma') \tag{6}$$

with

$$\alpha' = \rho^2_{p'o'} \tag{7a}$$

$$\beta' = \left(\rho_{p'o'} - \frac{\sigma_{p'}}{\sigma_{o'}}\right)^2 \tag{7b}$$

$$\gamma' = \frac{\left(\overline{p'} - \overline{o'}\right)^2}{\sigma^2_{o'}}. \tag{7c}$$

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 $\mathrm{MSE}_r$:

$$\mathrm{MSE}_r = \sigma^2_{o'}(1 + \epsilon') \tag{8}$$

where

$$\epsilon' = \frac{\overline{o'}^2}{\sigma^2_{o'}} \tag{9}$$

is non-zero if the map mean of the observations differs from the map mean of the reference prediction.

Finally, substitution of eqs. (6) and (8) in eq. (2) yields the Murphy–Epstein decomposition of the skill score (Murphy and Epstein, 1989):

$$\mathrm{MSESS} = \frac{\alpha' - \beta' - \gamma' + \epsilon'}{1 + \epsilon'}. \tag{10}$$

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

### 2.2. *Reference model of zero morphological change*

In morphodynamic modelling, the predictand is the bathymetry, such that $p$ and $o$ in eq. (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. (3) to (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) leads to a skill score valid for the zero change reference model:

$$\mathrm{MSESS}_{\mathrm{ini}} = 1 - \frac{\langle(\Delta z_p - \Delta z_o)^2\rangle}{\langle\Delta z_o^2\rangle} \tag{11}$$

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 1 shows the ranking proposed by Sutherland et al. (2004) for the skill formulation according to eq. (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).

Table 1: Classification according to Sutherland et al. (2004) for the MSE skill score as in eq. (11)

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

With the anomalies equal to the cumulative sedimentation/erosion fields, eqs. (7) and (9) can be written as $\alpha' = \rho^2_{\Delta z_p \Delta z_o}$, $\beta' = \left(\rho_{\Delta z_p \Delta z_o} - {}^{\sigma_{\Delta z_p}}/_{\sigma_{\Delta z_o}}\right)^2$, $\gamma' = {}^{\left(\overline{\Delta z_p} - \overline{\Delta z_o}\right)^2}/_{\sigma^2_{\Delta z_o}}$ and $\epsilon' = {}^{\overline{\Delta z_o}^2}/_{\sigma^2_{\Delta z_o}}$. For the normalization term $\epsilon'$, non-zero 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 non-zero $\gamma'$ 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, $\gamma'$ 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 $\beta'$ as measures of phase and amplitude errors, respectively, of the cumulative sedimentation/erosion fields (see section 2.1). Note that the phase and amplitude errors of predicted *bed levels* are given by $1 - \alpha$ and $\beta$ (eqs. (A.3a) and (A.3b)) rather than $1 - \alpha'$ and $\beta'$. 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 $\alpha'$ 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 A), a non-perfect 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 $\alpha'$. Therefore, we prefer to consider $\alpha'$ 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 $\alpha'$ 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 non-zero amplitude error $\beta'$ indicates that "the wrong *volumes* of sand have been moved", whereas Gerritsen et al. (2011) refer to $\beta'$ as a transport rate error.

Section 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.

### *2.3. Morphodynamic model skill as (mis)perceived using the zero change model*

In eq. (11), the MSE is normalized with $\mathrm{MSE}_r$ 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 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.

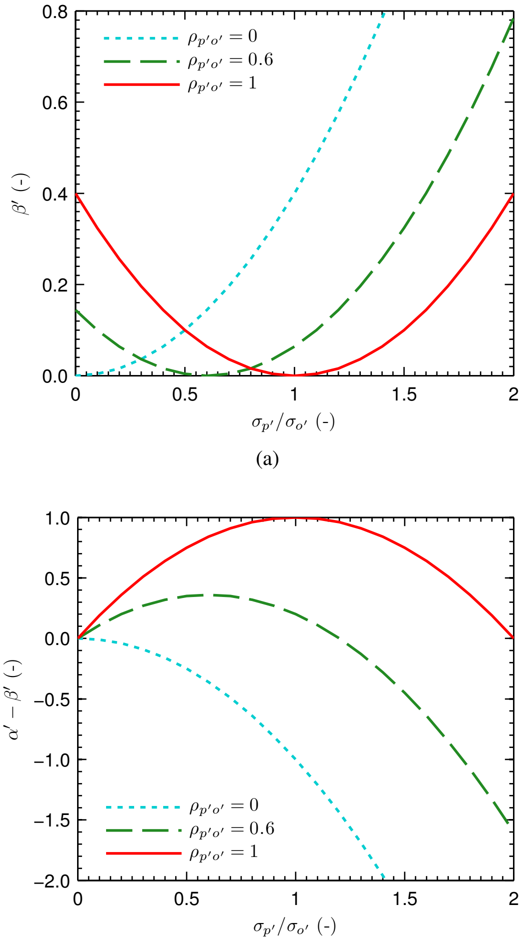
In conclusion, observed mean-squared cumulative bed changes cannot be expected to accurately reflect and thus effectively neutralize the level of difficulty among different prediction situations and times in a simulation. This places severe limits on the general validity of a comparative analysis based on $\mathrm{MSESS_{ini}}$. On a case-by-case basis, $\mathrm{MSESS_{ini}}$, notably its time-evolution for a trend, may still provide useful information. Therefore, section 3 thoroughly investigates how to interpret the temporal variation of $\mathrm{MSESS_{ini}}$ for a real-life case that shows a consistent bathymetric development away from the initial bed.

### *2.4. Underestimation of the variance of bed changes through the use of* $\mathrm{MSESS_{ini}}$

In this section, we demonstrate that $\mathrm{MSESS_{ini}}$ is prone to reward predictions that underestimate the overall magnitude of bed changes. To this end, we analyse the Murphy–Epstein decomposition of $\mathrm{MSESS_{ini}}$ and especially the amplitude error $\beta'$.

The behaviour of $\beta'$, which is controlled by $\sigma_{p'}/\sigma_{o'}$ and $\rho_{p'o'}$ (eq. (7b)), is shown in Figure 1a for $\rho_{p'o'} = 0, 0.6$ and 1. The line for $\rho_{p'o'} = 0.6$ is characteristic of the behaviour of $\beta'$ 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 $\beta'$ is non-zero. In fact, the amplitude error $\beta'$ is minimized for $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'}$. As a result, the interpretation of a non-zero $\beta'$ 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 $\mathrm{MSESS_{ini}}$ is maximized for $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'}$ (eq. (10) and fig. 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. (A.3b), this feature is inherited from the MSE, which is known for its tendency to reward the underestimation of



(a)

(b)

Figure 1: Amplitude error $\beta'$ and skill score MSESS$_{\text{ini}}$ = $\alpha' - \beta'$ (assuming $\gamma' = \epsilon' = 0$ in eq. (10)) versus $\sigma_{p'}/\sigma_{o'}$ for $\rho_{p'o'}$ equal to 0, 0.6 and 1: (a) $\beta'$ has a minimum for $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'}$, (b) the skill $\alpha' - \beta'$ is maximized for $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'}$.

the variability (e.g. Arpe et al., 1985; Gupta et al., 2009; Bosboom and Reniers, 2014a).

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-size 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 $\epsilon'$ are equal. From their fig. 4 and table 9, it can be seen that among the three predictions that have the same positive, but non-perfect 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 $\beta'$ is the smallest, viz. $\beta' = 0.01$). The values of $\gamma'$ 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. 1a[2]), is diagnosed with the highest skill (MSESS$_{\text{ini}}$ = 0.34, 0.29, 0.15 for $D_{50}$ = 0.5, 0.35, 0.25mm, 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 underestimated. Apparently, even when predictions are compared relative to the same initial bed, the characteristics of the MSESS$_{\text{ini}}$ 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 $\beta'$ is minimized and, unless compensated by systematic bias $\gamma'$, the MSESS$_{\text{ini}}$ 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. Clearly, these findings have implications for (automated) calibration as well as validation procedures that minimize MSE or MSESS$_{\text{ini}}$.

## 3. Illustration for the real-life case of Bornrif

In this section, the conventional validation method, discussed in section 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 MSESS$_{\text{ini}}$ 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, section 3.1 briefly describes the available observations and model set-up. Next, sections 3.2 and 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 section 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 MSESS$_{\text{ini}}$, is examined in section 3.5.

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

For the present validation, yearly bathymetric data up to depths of about 16m (JARKUS data; Minneboo, 1995) are available, interpolated to a 20×20m 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 20×20m grid that

[2] For $D_{50}$ = 0.5mm, 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 $D_{50}$ = 0.35mm and 0.25mm, $\sigma_{p'}/\sigma_{o'}$ = 0.88 and 1.07, respectively.

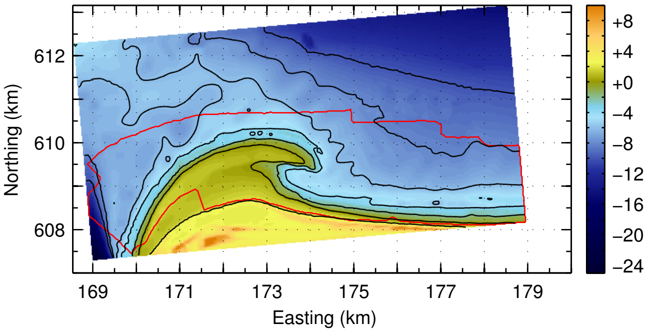


Figure 2: Initial bathymetry for the Bornrif simulation (1993) with the red polygon indicating the analysis region.

the JARKUS data were presented on. To that end, the computations were interpolated onto the observational grid. The red polygon in fig. 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 section 3.3).

### 3.2. Visual validation

The bathymetries and the yearly and cumulative sedimentation/erosion fields within the bounding polygon are shown in figs. 3 to 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. 3). From about 1998, migrating sand bars are observed at water depths larger than 5m 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 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. 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 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 the end of the simulation.

The cumulative bed changes (fig. 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. 3) deteriorates with time, it is much harder to visually judge the quality of the cumulative sedimentation/erosion fields over time (fig. 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 section 3.4 by comparison with the conventional error statistics that are discussed in the next section.

### 3.3. Conventional error statistics

The skill score $\text{MSESS}_{\text{ini}}$ according to eq. (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. 6a). According to table 1, the score qualifies as ‘good’ for all years. Based on $\text{MSESS}_{\text{ini}}$, 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. 6b)[3]. That nonetheless the skill, viz. the relative accuracy, increases with time is due to $\text{MSE}_{\text{ini}}$, the MSE of the reference prediction, increasing with time and, until 2002, at a faster rate than the MSE of the predictions (fig. 6b). With $\text{MSE}_{\text{ini}} = \langle \Delta z_o^2 \rangle$, 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 6a and 6b exemplify that, for a trend, the accuracy required for a certain level of skill decreases further into the simulation (section 2.3). In order to better value $\text{MSESS}_{\text{ini}}$ 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. (6) and (7) and eqs. (8) and (9), with $p' = \Delta z_p$ and $o' = \Delta z_o$, are shown in fig. 6c and fig. 6d, respectively. The MSE normalized with the variance of the observed anomalies, shown in fig. 6c, is dominated by the phase error $1 - \alpha'$ of the anomalies. The normalized sediment budget error $\gamma'$ decreases with time and only plays a role in the first half of the simulation, while the amplitude error $\beta'$ is negligible throughout the simulation. Figure 6d illuminates that the bias part $\epsilon' \sigma_{o'}^2$ of $\text{MSE}_{\text{ini}}$ is negligible ($\epsilon' \ll 1$), such that $\text{MSE}_{\text{ini}} \approx \sigma_{o'}^2$. The skill score (eq. (10)) is thus given by $\text{MSESS}_{\text{ini}} \approx 1 - {}^{\text{MSE}}/_{\sigma_{o'}^2} \approx \alpha' - \gamma'$ and from, say, 1999, $\text{MSESS}_{\text{ini}} \approx \alpha'$. Thus, the decrease of both the phase error $1 - \alpha'$ and the sediment budget error $\gamma'$ contributes to the increase 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 $\gamma'$ normalizes an absolute map-mean error $\text{MSE}_{\text{bias}} = (\overline{p} - \overline{o})^2 = \left(\overline{p'} - \overline{o'}\right)^2$ with the variance of the cumulative observed bed changes $\sigma_{o'}^2$ (eq. (7c)). Analysis showed

[3] 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.

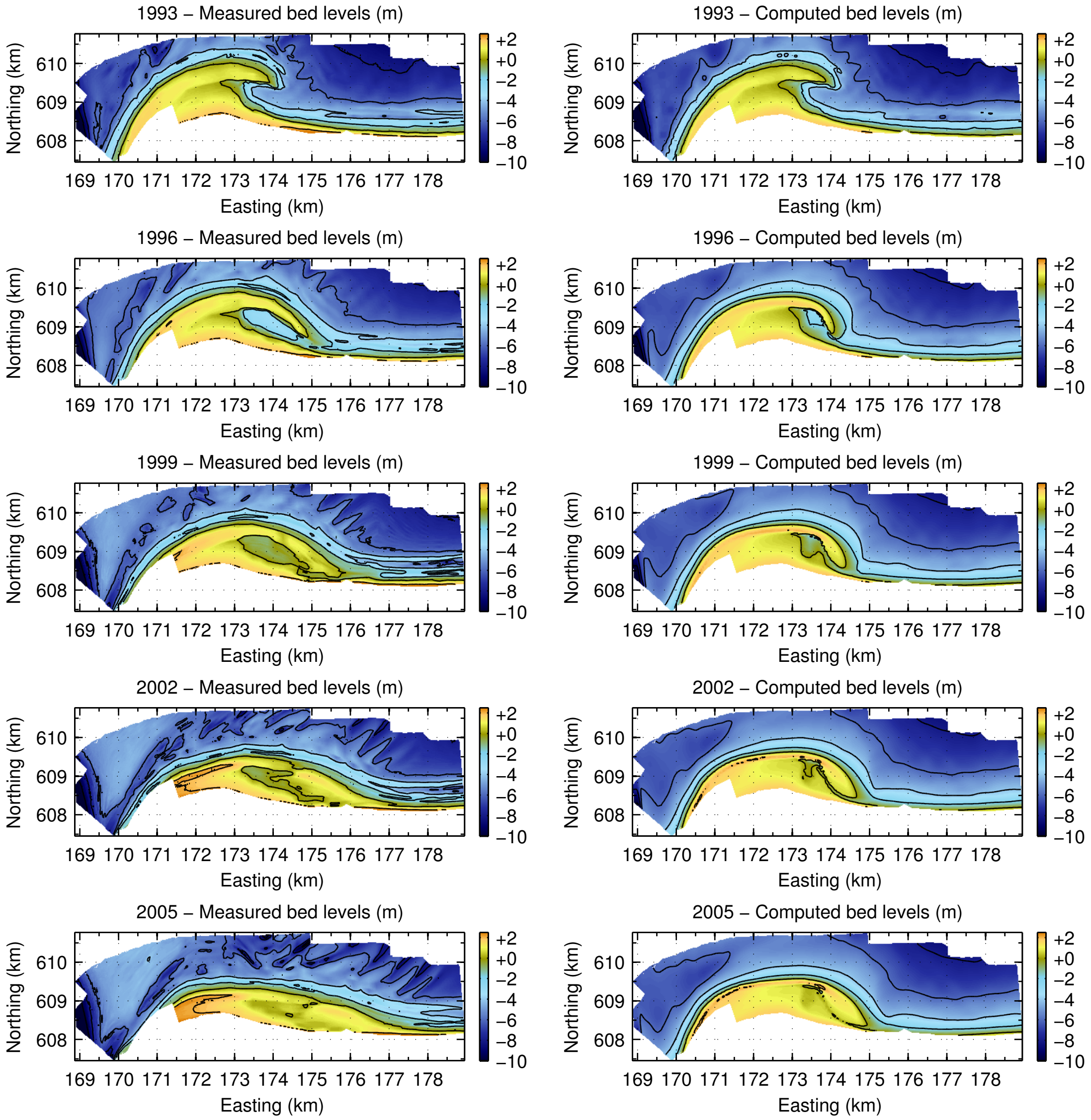

1993 – Measured bed levels (m)
1993 – Computed bed levels (m)
1996 – Measured bed levels (m)
1996 – Computed bed levels (m)
1999 – Measured bed levels (m)
1999 – Computed bed levels (m)
2002 – Measured bed levels (m)
2002 – Computed bed levels (m)
2005 – Measured bed levels (m)
2005 – Computed bed levels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

Figure 3: Measured (left) and computed (right) Bornrif bathymetries for the years 1993, 1996, 1999, 2002 and 2005 for the analysis region.

that the rapid decrease of $\gamma'$ 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 $\beta'$ (eq. (7b)) is the direct result of $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$ being relatively close together in value (fig. 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. 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 section 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. 7a).

The effect on the skill score is visualized in fig. 7b, which shows the behaviour of MSESS$_{\text{ini}}$ $= \alpha' - \beta'$ (eq. (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 $\beta'$ 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 (section 2.4).

With $\beta'$ negligible and $\gamma'$ vanishing after the first years of the simulation, the skill score MSESS$_{\text{ini}}$ peaks simultaneously with the phase association $\alpha'$ and the maximum value of MSESS$_{\text{ini}}$, in 2002, is fully determined by $\alpha'$ (figs. 6a and 6c). In section 2.2, we interpreted $\alpha'$ 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), $\alpha'$ 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, MSE$_{\text{bias}}$) 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 MSESS$_{\text{ini}}$ is inherited from the use of the MSE as the accuracy measure (cf. section 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 $\alpha'$).

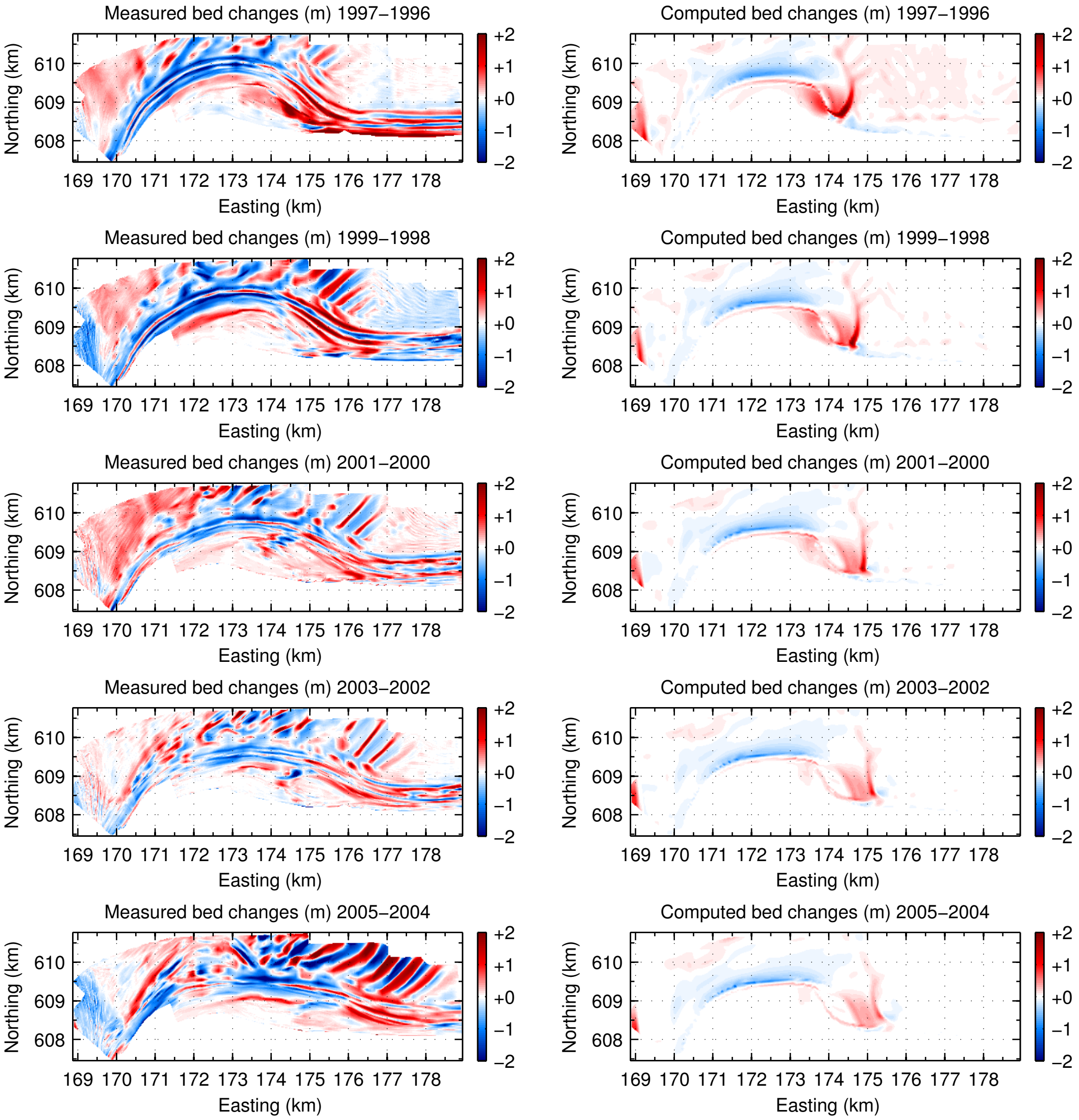

Measured bed changes (m) 1997−1996
Computed bed changes (m) 1997−1996
Measured bed changes (m) 1999−1998
Computed bed changes (m) 1999−1998
Measured bed changes (m) 2001−2000
Computed bed changes (m) 2001−2000
Measured bed changes (m) 2003−2002
Computed bed changes (m) 2003−2002
Measured bed changes (m) 2005−2004
Computed bed changes (m) 2005−2004
Northing (km)
608
609
610
Easting (km)
169 170 171 172 173 174 175 176 177 178
+2
+1
+0
−1
−2

Figure 4: Measured (left) and computed (right) Bornrif yearly bed changes for several years.

### 3.4. *Visual validation versus error statistics*

Sections 3.2 and 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. (10), MSESS$_{\text{ini}}$ 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 (section 3.2). A small bias, as in fig. 5, will most likely go unnoticed. Our impression, from fig. 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. 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. 5, whether the predictions in 1996 or 2002 are of higher quality. On the contrary, the development of MSESS$_{\text{ini}}$ over the course of the simulation was seen to merely report the correlation $\rho_{p'o'}$ between cumulative sedimentation/erosion fields (fig. 6c), such that the 2002 predictions are diagnosed with maximum skill (fig. 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 (section 3.2). Clearly, even though 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. (A.2) and fig. 8a), as opposed to of the anomalies (eq. (6) and fig. 6c). Although the variance of the observations $\sigma_o^2$ varies in time (fig. 8b), it is relatively constant as compared to $\sigma_{o'}^2$ (fig. 6d). Hence, where the MSE normalized with $\sigma_{o'}^2$ behaves quite differently from the MSE itself, the MSE normalized with $\sigma_o^2$ increases in time as the MSE does. From fig. 8a, $MSE/\sigma_o^2$ can be seen to be dominated by the phase error $1-\alpha$, which increases with time as a result of the decreasing correlation $\rho_{po}$ between predicted and observed bed levels (fig. 8c). Analogously, the most obvious finding from the visual validation of bed levels (fig. 3) was the decreasing overall agreement in structural similarity between the measured and predicted bathymetric fields. The slight increase in amplitude error $\beta$ is governed by the fact that $\rho_{po}$ decreases faster with time than $\sigma_p/\sigma_o$ (fig. 8c). Note further that,

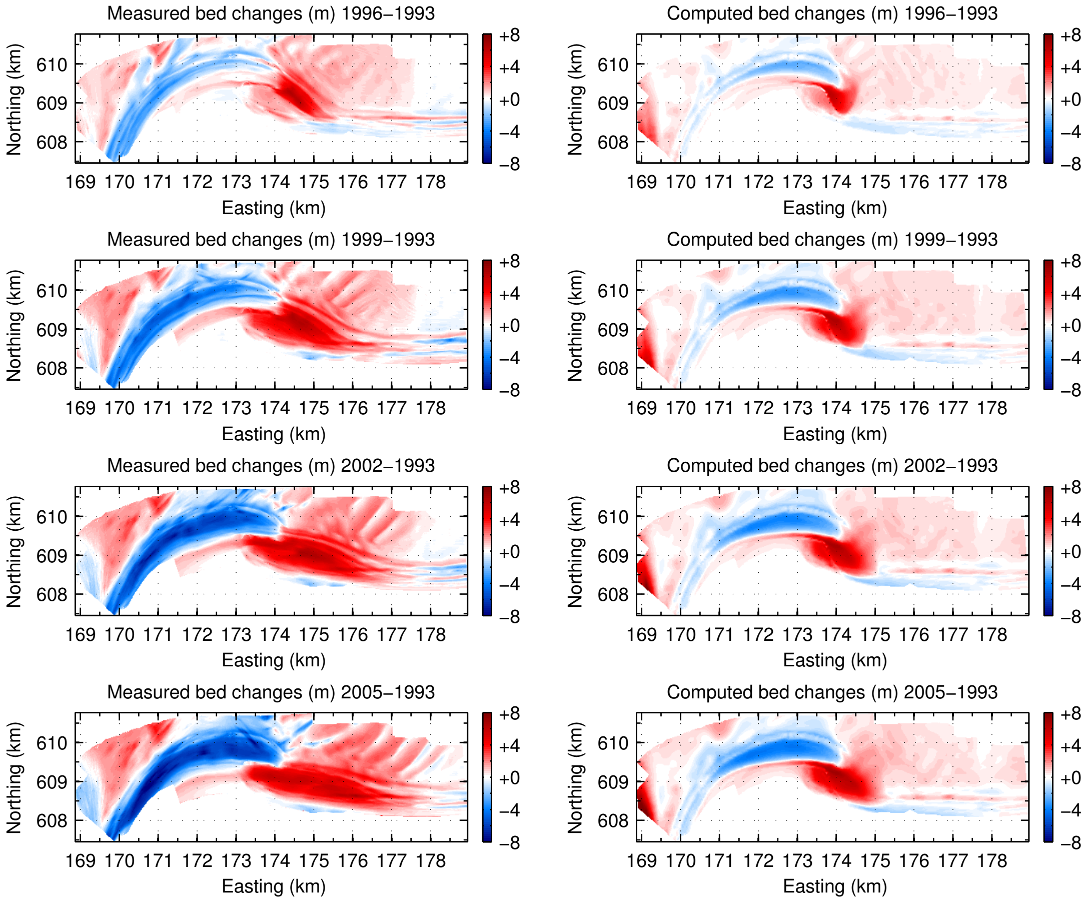

Measured bed changes (m) 1996−1993
Computed bed changes (m) 1996−1993
Measured bed changes (m) 1999−1993
Computed bed changes (m) 1999−1993
Measured bed changes (m) 2002−1993
Computed bed changes (m) 2002−1993
Measured bed changes (m) 2005−1993
Computed bed changes (m) 2005−1993
Northing (km)
610
609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

Figure 5: Measured (left) and computed (right) Bornrif cumulative bed changes for 1996, 1999, 2002 and 2005 with respect to the initial bed of 1993.

analogously to $\beta'$, $\beta$ would have been larger, if only slightly, for $\sigma_p/\sigma_o = 1$.

The normalized metrics for the bed levels, $\rho_{po}$ and $\sigma_p/\sigma_o$, provide information not contained in the anomalies. For instance, from fig. 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 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. 3).

In conclusion, 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. 3 and 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$_{\text{bias}}$, $\sigma_{p'}/\sigma_{o'}$, $\rho_{p'o'}$, $\sigma_p/\sigma_o$ and $\rho_{po}$ respectively.

### *3.5. The effect of various spatial scales*

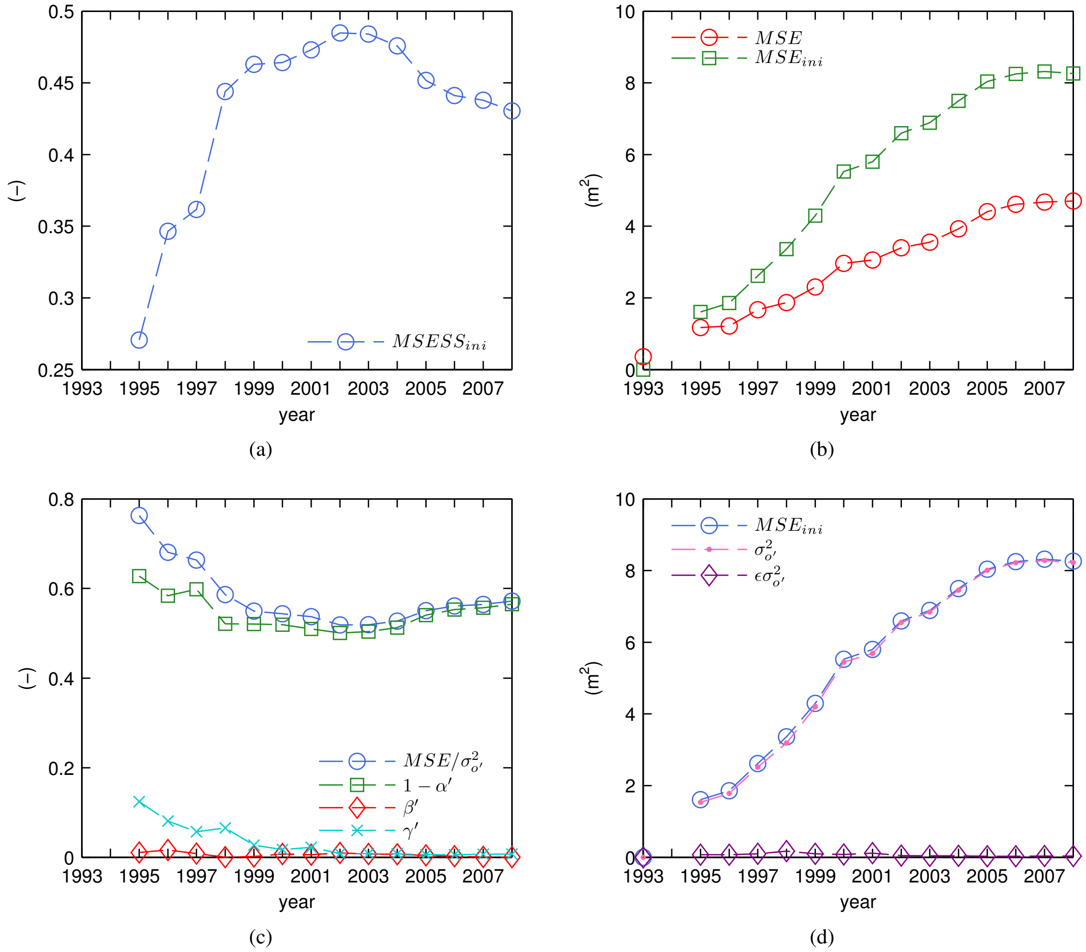
The various statistics, discussed in section 3.4, inevitably combine information across a range of spatial scales. Hence, it is nontrivial to relate $\rho_{po}$ and $\sigma_p/\sigma_o$ or $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$ to particular features of interest in the morphology or the fields of cumulative change, respectively. The range over which spatial scales are lumped together is especially wide for the normalized bed level metrics, $\rho_{po}$ and $\sigma_p/\sigma_o$, in which scales up to the size of the model domain play a role (cf. section 3.4). By implication, the values of $\rho_{po}$ or $\sigma_p/\sigma_o$ 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 (section 3.2). This suggests that the relative contribution of the smaller scales to $\rho_{p'o'}$, $\sigma_{p'}/\sigma_{o'}$ and 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. (3) and $r = 0$ in eq. (5). Upon substitution, eq. (2) yields:

$$\text{MSESS}_{\Delta z,1} = 1 - \frac{\langle(\Delta z_{p,1} - \Delta z_{o,1})^2\rangle}{\langle \Delta z_{o,1}^2 \rangle}. \tag{12}$$

Note that if the period of bed changes was taken as the simulation duration up to the evaluation time, we obtain MSESS$_{\text{ini}}$ (eq. (11)). For all years of the Bornrif simulation, the relative accuracy of



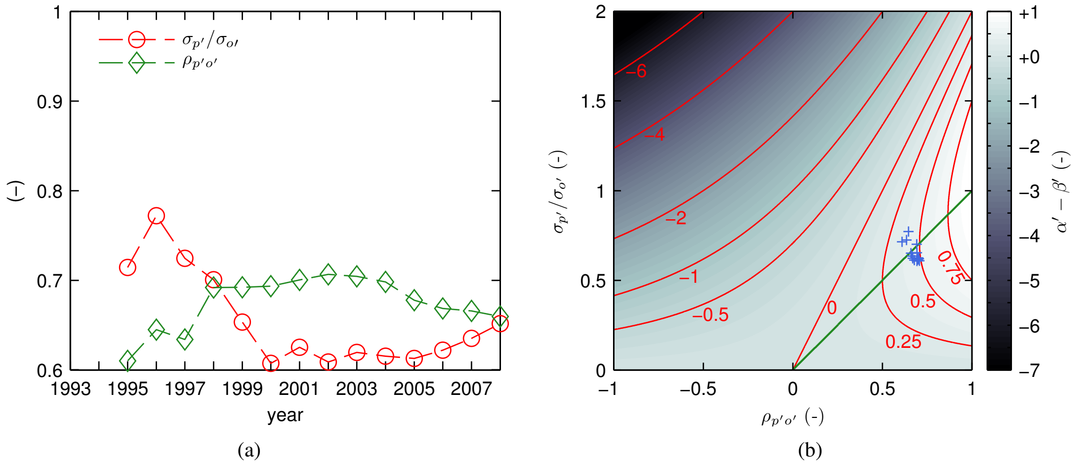
(a)

(b)

(c)

(d)

Figure 6: Model performance for Bornrif: (a) MSE skill score with the zero change model as the reference, $\text{MSESS}_{\text{ini}}$, (b) MSE of the computations and $\text{MSE}_{\text{ini}}$ of the initial bed (zero change reference model), (c) MSE normalized with the variance of the cumulative observed bed changes and its decomposition, eqs. (6) and (7) and (d) $\text{MSE}_{\text{ini}}$ and its decomposition, eqs. (8) and (9).



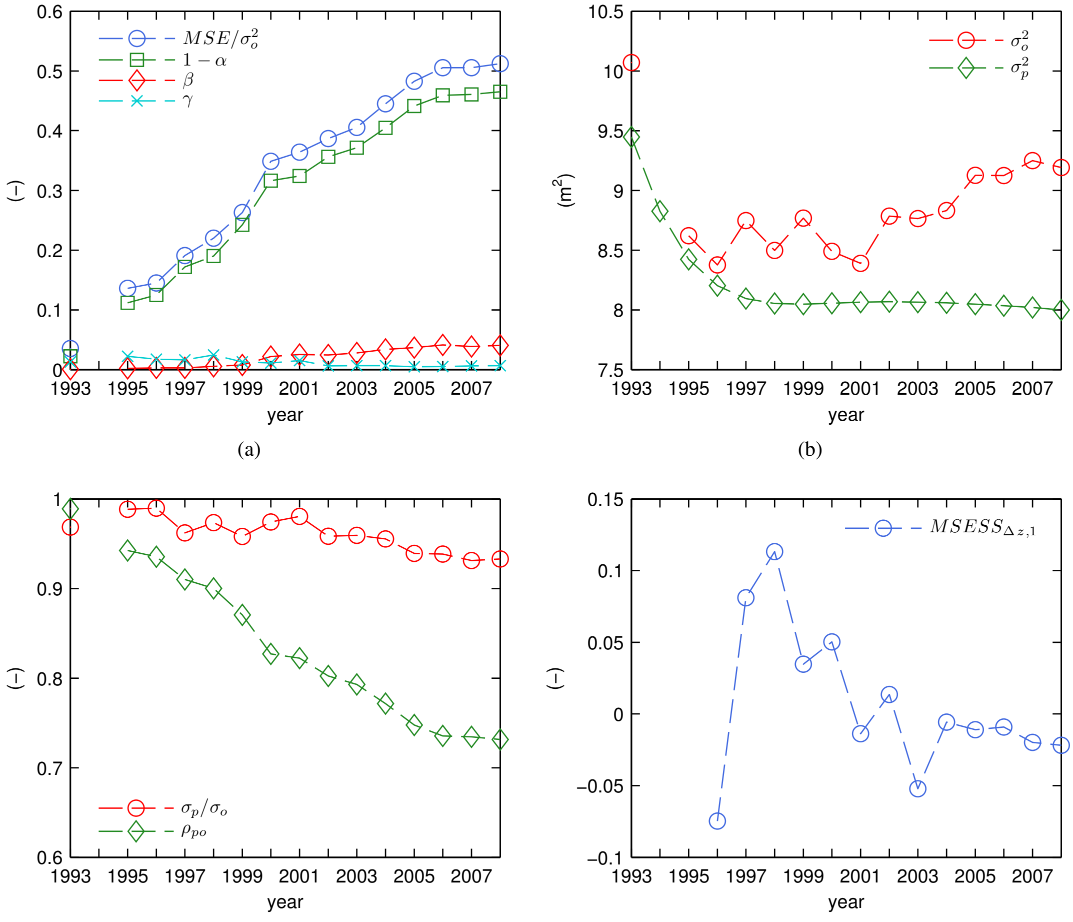
(a)

(b)

Figure 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 $\text{MSESS}_{\text{ini}} \approx \alpha'$.

yearly change, $\text{MSESS}_{\Delta z,1}$, is low or negative (fig. 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. 6d). Since eq. (12) does not consider any cumulative effect on time-scales 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 $\text{MSESS}_{\text{ini}}$ at the beginning of the Bornrif simulation (fig. 6a) are mainly due to unskilful smaller spatial scales. When, over time, the relative contribution of these smaller scales to the cumulative



(a)

(b)

(c)

(d)

Figure 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$^3$, the 1993 measured and computed parameters differ slightly.

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 to develop validation methods that distinguish between various spatial scales.

## 4. Summary and discussion

The use of MSESS$_{\text{ini}}$ (eq. (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, section 4.1 focuses on the effect of the choice of the zero change reference model. Second, section 4.2 summarizes the aspects of model performance captured by MSESS$_{\text{ini}}$ as well as by visual validation.

### *4.1. The zero point at the scale of skill*

The MSESS$_{\text{ini}}$ is frequently used to compare morphodynamic model performance across different prediction situations. We have demonstrated however, that the validity of the ranking based on MSESS$_{\text{ini}}$ (table 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 MSESS$_{\text{ini}}$ 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 (section 2.3) showed that this assumption cannot be expected to hold.

In connection with the above, it was argued that MSESS$_{\text{ini}}$ may also misreport the temporal evolution of model skill. For interseasonal 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 (section 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 (section 2.3). If the accuracy of the reference model decreases in time at a faster rate than the accuracy of the predictions, the MSESS$_{\text{ini}}$ may even increase with time, while the agreement between modelled and observed bathymetry strongly decreases, as was seen for the Bornrif (section 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 (section 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 non-trivial. 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 section 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 MSESS$_{\text{ini}}$, 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 MSESS$_{\text{ini}}$ alone.

### *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 (section 3.2). These notions can be quantified by e.g. $\sigma_{p'}/\sigma_{o'}$, $\rho_{p'o'}$, $\sigma_p/\sigma_o$, $\rho_{po}$, MSE and MSE$_{\text{bias}}$, respectively (sections 3.3 and 3.4). Summary metrics, such as the MSE and the MSESS$_{\text{ini}}$, were seen to provide an implicit weighting of systematic bias terms as well $\rho_{po}$ and $\sigma_p/\sigma_o$ and $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$, respectively. Unfortunately, in doing so, MSE and MSESS$_{\text{ini}}$ tend to reward the underprediction of the variance of bed levels and bed changes, respectively, as shown in section 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, $1 -$ MSESS$_{\text{ini}}$ 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 MSESS$_{\text{ini}}$ is maximized for $\sigma_{p'}/\sigma_{o'} = \rho_{p'o'}$, hence for too small overall bed changes (section 2.4 and fig. 7b). For a real-life case, taken from literature, this was shown to have resulted in the ranking of predictions based on MSESS$_{\text{ini}}$ being inconsistent with expert judgement (section 2.4). Similarly, since for the Bornrif simulation $\rho_{p'o'}$ and $\sigma_{p'}/\sigma_{o'}$ are close together in value (fig. 7a), the skill levels are dominated by $\rho_{p'o'}$. As a result, the development of MSESS$_{\text{ini}}$ in time was seen to merely report the correlation $\rho_{p'o'}$ between cumulative sedimentation/erosion fields (section 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 MSESS$_{\text{ini}}$; 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 MSESS$_{\text{ini}}$.

In morphodynamic model validation, the MSESS$_{\text{ini}}$ is sometimes supplemented with its Murphy–Epstein decomposition (eq. (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 $\beta'$, respectively, are not necessarily in line with the morphologists' intuitive definition. The phase association $\alpha'$ (eq. (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 $\alpha'$ does not distinguish between positive and negative correlations). Further, when neglecting systematic bias, $\sigma_{p'}/\sigma_{o'}$ rather than $\beta'$ (eq. (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 $\gamma'$ (eq. (7c)) is also non-trivial, 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 $\gamma'$, 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 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 (section 3.5).

In summary, although frequently used as the main indicator of morphodynamic model skill, the use of 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 (sections 3.3 and 3.4). In doing so, it is crucial, yet non-trivial, 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, 2014b). 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).

## 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, 2014a). 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, 2014b). 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.

## Appendix A. Murphy–Epstein decomposition of MSE

Algebraic manipulation of the MSE, eq. (3), leads to (Murphy, 1988):

$$\text{MSE} = \sigma_p^2 + \sigma_o^2 - 2\sigma_p\sigma_o\rho_{po} + (\overline{p} - \overline{o})^2 \tag{A.1}$$

where $\overline{p}$ and $\overline{o}$ are the weighted map means and $\sigma_p$ and $\sigma_o$ the weighted standard deviations of the predictions $p$ and the observations $o$, respectively, and $\rho_{po}$ 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 $\sigma_{po}$ 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. (A.1) to arrive at (Murphy and Epstein, 1989):

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

where

$$\alpha = \rho_{po}^2 \tag{A.3a}$$

$$\beta = \left(\rho_{po} - \frac{\sigma_p}{\sigma_o}\right)^2 \tag{A.3b}$$

$$\gamma = \frac{(\overline{p} - \overline{o})^2}{\sigma_o^2}. \tag{A.3c}$$

Here, $\gamma$ is a normalized map-mean error. The term $\beta$ is the conditional bias, which is non-zero 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 $\alpha$ 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. (A.1) to (A.3) are equally valid when $p$ and $o$ are replaced with $p'$ and $o'$, respectively.

## References

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, B.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.

Bosboom, J., Reniers, A.J.H.M., 2014a. 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., 2014b. Scale-selective validation of morphodynamic models, in: Proceedings of the 34th International Conference on Coastal Engineering, Seoul, South-Korea.

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.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Monthly Weather Review 78, 1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brier, G.W., Allen, R.A., 1951. Verification of weather forecasts. Compendium of Meteorology , 841–848.

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

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

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.

Gallagher, E.L., Elgar, S., Guza, R., 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 1, 166–172.

Gupta, H., Kling, H., Yilmaz, K., Martinez, G., 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.

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.

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.

Lesser, G., Roelvink, J., van Kester, J., 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.

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.

McCall, R., de Vries, J.V.T., Plant, N., Dongeren, A.V., Roelvink, J., Thompson, D., Reniers, A., 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., 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).

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.

Orzech, M.D., Reniers, A.J., 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.

Roelvink, D., Reniers, A., 2012. A guide to Modeling Coastal Morphology. volume 12. World Scientific Publishing Company.

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.

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

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

Ruggiero, P., Walstra, D., 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.

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

Stive, M.J., de Schipper, M.A., Luijendijk, A.P., Aarninkhof, S.G., van Gelder-Maas, C., van Thiel de Vries, J.S., 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., Soulsby, R., 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.

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.

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

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. 3 ed., Academic Press.

Williams, J.J., de Alegra-Arzaburu, A.R., McCall, R.T., Dongeren, A.V., 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.