Coastal Engineering 94 (2014) 112–125

Contents lists available at ScienceDirect

Coastal Engineering

journal homepage: www.elsevier.com/locate/coastaleng

# On the perception of morphodynamic model skill

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

[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

ARTICLE INFO

*Article history:*
Received 7 February 2014
Received in revised form 15 August 2014
Accepted 18 August 2014

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

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 favor 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 judgment of experts. Guidelines are suggested for how to adjust calibration and validation procedures to be more in line with a morphologist's expert judgment.



## 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 modelers (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 modelers 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

* Corresponding author. Tel.: +31 15 27 84606; fax: +31 15 27 85124.
*E-mail 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).

http://dx.doi.org/10.1016/j.coastaleng.2014.08.008
0378-3839/

standard in quantitative judgment 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 behavior 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 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 unskillful, 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, which is a 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 judgment. 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:

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

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:

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

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

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

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