SCALE-SELECTIVE VALIDATION OF MORPHODYNAMIC MODELS

J. Bosboom[1] and A.J.H.M. Reniers[1,2]

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

Keywords: morphodynamic modelling; model validation; spatial scales; local statistics; skill

INTRODUCTION

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

In other fields, notably meteorology, scale-dependent verification methods have been proposed that are able to describe the scale at which a forecast attains a particular level of skill (e.g. Roberts and Lean 2008); for an overview, see Gilleland et al. (2010). Also, in the field of image processing, Wang et al. (2004) determine the closeness of images using a multi-scale method, which incorporates image details at different resolutions. These methods typically utilize band-pass filters (Fourier, wavelets, etc.) or smoothing filters for the separation of scales. For 2D morphology and arbitrarily shaped model domains, the application of such band-pass filters and the physical interpretation of the results is far from trivial. Methods based on smoothing filters are appealing due to their simplicity, but often limited in the aspects of model performance that can be considered. For instance, no information on spatial variation of skill in the model domain is provided.

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

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

[1] Department of Hydraulic Engineering, Faculty of Civil Engineering and Geosciences, Delft University of Technology, P.O.Box 5048, 2600 GA Delft, The Netherlands. Corresponding author: j.bosboom@tudelft.nl
[2] Deltares, Unit Marine and Coastal Systems, P.O.Box 177, 2600 MH Delft, The Netherlands

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

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

SCALED SKILL

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

Aspects of model performance: structural and amplitude similarity per scale

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

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

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

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

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

which, with q =1, is the form as used by Wang et al. (2004) and Koh et al. (2012) and named contrast measure and variance similarity, respectively.

In morphodynamic modelling, where the predictand is the bathymetry, the interpretation of ρ_{po} and $\hat{\sigma} = \sigma_p / \sigma_o$ in terms of bed features is far from trivial, since multiple scales are generally present in the observed and computed bathymetry (Fig. 1) and larger scales may overwhelm the smaller scales. Fig. 1 (middle panel) indicates that the overall correlation can be negative, whilst the correlation can be positive if we zoom in to a smaller area. This situation can of course also be reversed, with positive correlation for larger scales and negative correlation for smaller scales (Fig. 1, top panel). The latter situation may be closer to what we expect from a typical morphodynamic simulation. Not only the correlation but also the ratio of the standard deviations between predictions and observations may vary

with spatial scale. For example, Fig. 1 (bottom panel) shows an overestimation of the variability for the larger scale and an underestimation for the smaller scale.

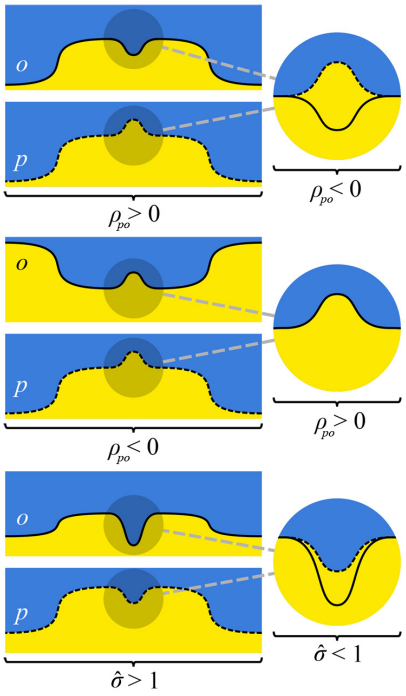

Figure 1. Scale-dependency of comparisons between observations *o* and predictions *p*. Top panel: the correlation is higher at the larger scale; middle panel: the correlation is higher at the smaller scale. Bottom panel: the amplitude similarity is also dependent on the scale.

Localized statistics

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

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

$$\overline{o}_i = \sum_j w_{ij} o_j \tag{3}$$

$$\overline{p}_i = \sum_j w_{ij} p_j \tag{4}$$

with w_{ij} is a weighting factor dependent on the proximity to the location *i* and $\sum_j w_{ij} = 1$. All results shown in this paper are obtained with a very simple (and fast) window, viz. a rectangular window with a width *W,* uniform weights within the window and $w_{ij} = 0$ elsewhere in the domain (Fig. 2). Hence, $w_{ij} = w_{ij}(W)$. A more sophisticated approach uses a distance decay function given by for instance a bi-square kernel with a variable bandwidth (see e.g. Fotheringham et al. 2002).

Of course, Eqs. 3 and 4 simply compute a (weighted) moving average. However, we can now extend the concept to arbitrary statistics, for instance the standard deviations $\sigma_{o,i}$ and $\sigma_{p,i}$ of observations and predictions, respectively:

$$\sigma_{o,i} = \left[\sum_j w_{ij} \left(o_j - \overline{o}_i \right)^2 \right]^{1/2} \tag{5}$$

$$\sigma_{p,i} = \left[\sum_j w_{ij} \left(p_j - \overline{p}_i \right)^2 \right]^{1/2} \tag{6}$$

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

$$\rho_{po,i} = \left(\sigma_{o,i}\sigma_{p,i}\right)^{-1} \sum_j w_{ij}\left(o_j - \overline{o}_i\right)\left(p_j - \overline{p}_i\right) \;,\quad -1 \le \rho_{po,i} \le 1 \qquad (7)$$

Note that in Eqs. 6 and 7, the local rather than the global mean values are used.

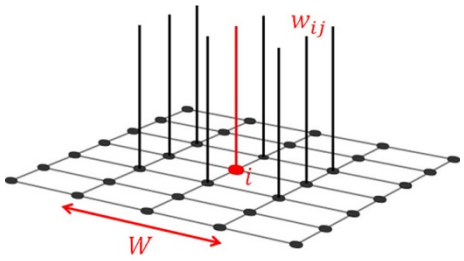

Figure 2. The rectangular window, around grid-point *i*, with window width *W* and weights w_{ij}.

Now, the local amplitude similarity is given by:

$$\eta_i = \left(\frac{2}{\hat{\sigma}_i + \hat{\sigma}_i^{-1}}\right)^q \quad \text{with} \quad \hat{\sigma}_i = \sigma_{p,i}/\sigma_{o,i} \;,\; 0 \le \eta_i \le 1 \qquad (8)$$

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

How to construct a skill score?

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

$$\mathrm{MSE}_{\mathrm{fluct}} = \sigma_o^2\left[1 - \rho_{po}^2 + \left(\rho_{po} - \hat{\sigma}\right)^2\right] \qquad (9)$$

Between two predictions with the same positive correlation, $\mathrm{MSE}_{\mathrm{fluct}}$ is minimized for $\hat{\sigma} = \rho_{po}$, hence for $\sigma_p = \rho_{po}\sigma_o$. In the case of a negative correlation, $\mathrm{MSE}_{\mathrm{fluct}}$ is minimized for $\hat{\sigma} = 0$ and thus for $\sigma_p = 0$. As a consequence, the MSE tends to reward the underestimation of the variability (see also Bosboom et al. in press). Nonetheless, a morphologist may prefer features to be predicted at the right amplitude albeit displaced above a featureless prediction (Bosboom and Reniers 2014). Therefore, we use an alternative weighting with the following behaviour: for any given variance the skill score increases monotonically with increasing correlation and for any given correlation the skill score increases as the modelled variance approaches the observed variance (Taylor, 2001).

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

$$S_i = \tfrac{1}{2}\left(1 + \rho_{po,i}\right)^m \eta_i^n \;,\quad 0 \le S_i \le 1 \qquad (10)$$

Note that S_i is a function of the window width W. The weighting of structural and amplitude similarity must, to a certain extent, be decided upon subjectively. The coefficients m and n allow the user to define the most appropriate weighting for the situation under consideration. In this paper, we have used $m = 1$ and $n = 1$ in Eq. 10 and $q = 2$ in Eq. 8.

A domain-averaged skill score S as a function of W can be obtained by averaging S_i (Eq. 10) over all grid-points i. We hypothesize that the smaller scales, down to the grid scale, are not as well predicted as the larger scales up to the scale of the entire domain, and that there is a minimum spatial scale above which the skill is sufficient, i.e. larger than a user-defined target skill (Fig. 3). For a real-life case, this hypothesis is put to the test in the next section.

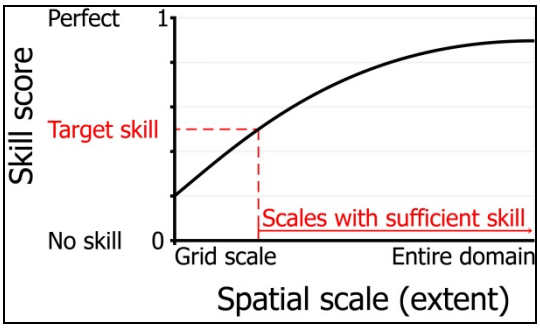

Figure 3. Hypothesized qualitative behaviour of the skill score *S* versus the spatial scale, which ranges from the grid scale to the entire domain. For larger spatial scales the skill value approaches the whole-map skill value $S = 1/2(1+\rho_{po})\eta$ computed using the values at all grid-points.

EXAMPLE

In this section, we demonstrate our method by applying it to measured and computed bathymetric fields for the Bornrif, a dynamic attached bar at the North-western edge of the Wadden Sea island of Ameland. First, we briefly describe the measurements and computations. Next, we show the maps of local statistics that are subsequently pooled into map-mean values per spatial scale. Finally, we explore the relationship between information richness and skill.

Bornrif

The Bornrif morphodynamic evolution was computed with Delft 3D from 1993 to 2008, using a grid with a resolution of 50 × 50 m in the central part of the model domain and 100 × 50 m closer to the model boundaries (Achete et al. 2011). A detailed description of this Delft3D simulation and the available data is found in Bosboom et al. (in press). Here, we focus on the results for 1998, hence five years after the start of the simulation (Fig. 4).

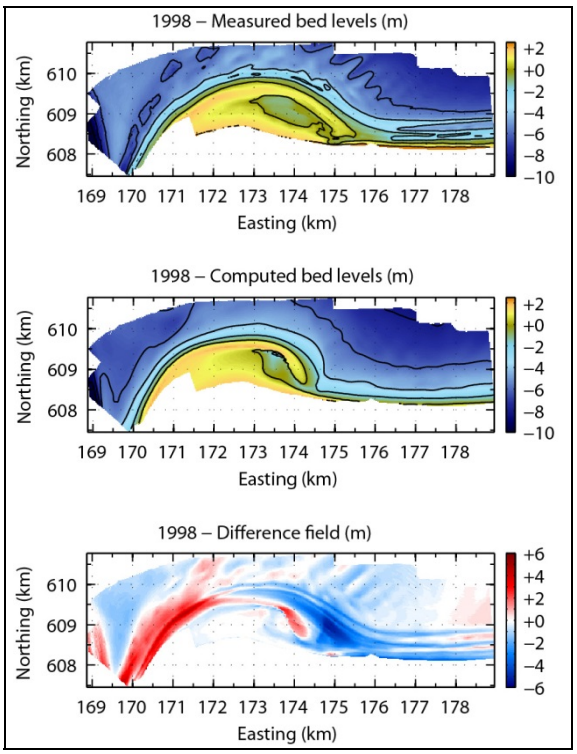

Figure 4. Measured (top panel) and computed (middle panel) Bornrif bathymetries for 1998 and the difference field *p-o* between predictions *p* and observations *o* (lower panel).

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

The analysis region, as shown in Fig. 4, covers only that part of the computational domain for which data are available during the entire simulation duration. In order to retain all observed scales, the spatial validation analysis is performed on the 20 × 20 m grid that the data were presented on. To that end, the computations were first interpolated onto the observational grid. In the following we demonstrate typical results of applying the method of scaled skill. The central validation question is: how skilful is the model in the various regions and at the various spatial scales that can be discerned?

Maps of local statistics

Areal maps of structural similarity $\rho_{po,i}$, amplitude similarity $\hat{\sigma}_i$ and η_i, and pattern skill S_i provide information on local differences in quality (Fig. 5). Such maps can be produced for various spatial scales (i.e. areal sizes of focus). Fig. 5 shows the results at three window sizes W = 0.16, 0.4 and 0.8km. There is a wealth of information in these figures; here we will only point out some main aspects.

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

On the contrary, there are also small-scale patches of negative correlation that are not present anymore at the larger scales, for instance in regions further offshore and in the nearshore region east of the Bornrif. In these areas, a low structural similarity $\rho_{po,i}$ is combined with a low amplitude similarity η_i, which can be seen - from $\hat{\sigma}_i$ being close to zero - to be due to an underestimation of the variability. This indicates small-scale, observed features that are not reproduced in the predictions, namely the sand bars at deeper water and the nearshore bars.

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

Pooled skill scores

Another way of looking at the quality variation is by making histograms of the quality maps (Fig. 6). The first column clearly shows that grid-points with negative correlation at small spatial scales obtain a positive correlation at larger scales. A similar trend can be observed from the second column that shows the amplitude similarity. The third column shows that, as a result, the percentage of the model domain with low pattern skill scores decreases with spatial scale, as was apparent from the pattern skill maps as well (Fig. 5).

The red lines in Fig. 6 show the domain-averaged values of the quality metrics for the three window sizes that are considered. Not surprisingly given the above, the quality according to each of these metrics increases with spatial scale. Apparently, the Bornrif morphology can be thought to consist of smaller-scale features that are not well represented by the model, on top of a larger-scale morphology that is better predicted.

When extending this analysis to a range of window sizes, we obtain Fig. 7, which shows the structural similarity, amplitude similarity and pattern skill versus window size. At the scale of the entire domain, the skill is very high, since the larger-scale morphology is reasonably well represented. However, at the smaller scales of the spit and the sand bars the skill is lower. Based on this figure, we can determine the smallest useful scale, viz. the smallest areal size with a certain desired level of skill. If the target skill is set to for instance 0.7 (as in Fig. 3), the smallest scale with sufficient skill is about 0.6km.

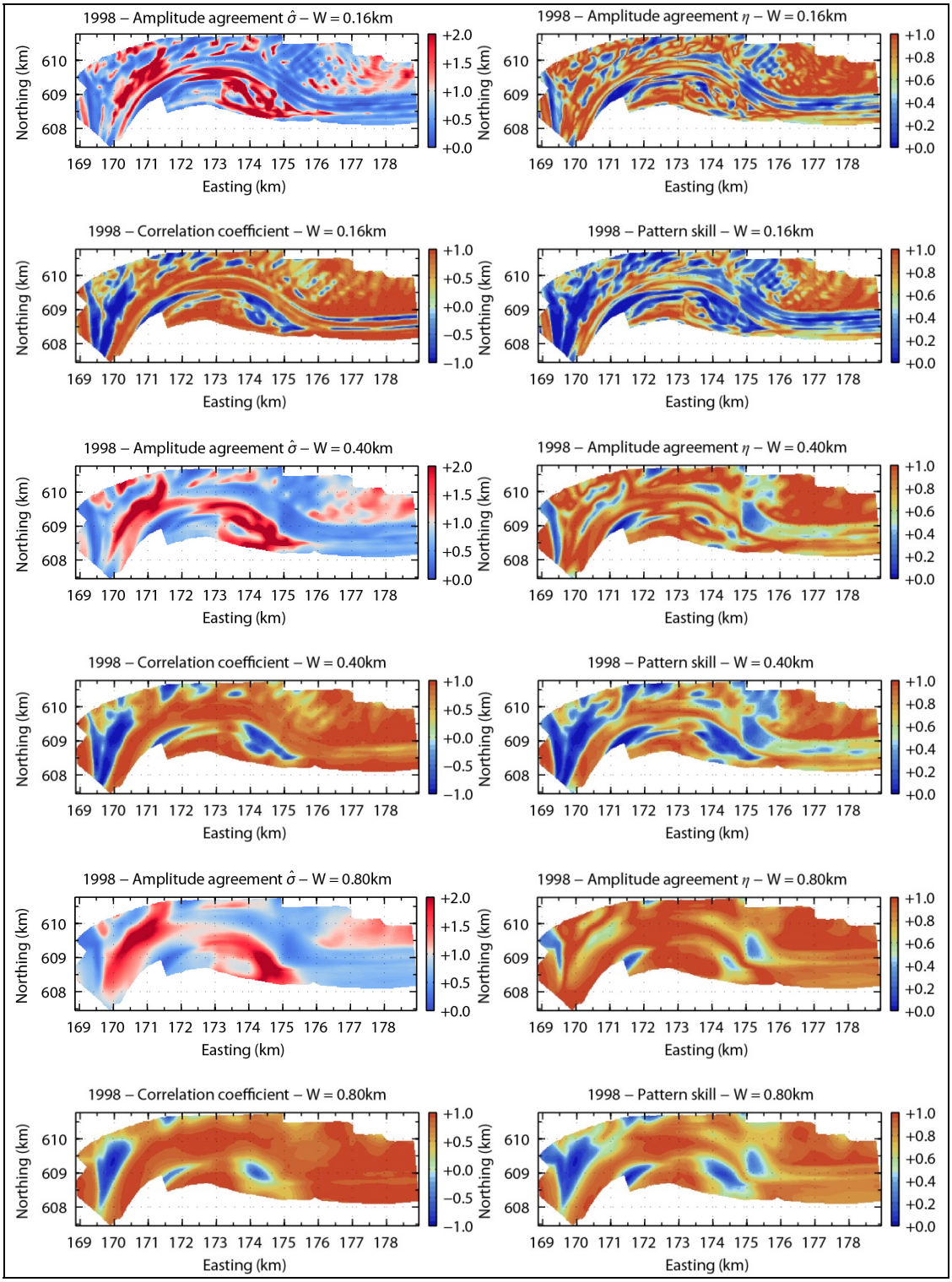

Figure 5. Normalized maps of structural and amplitude similarity and pattern skill for three different window sizes *W* = 0.16, 0.4 and 0.8km. For all quality metrics a value of 1 represents perfect agreement.

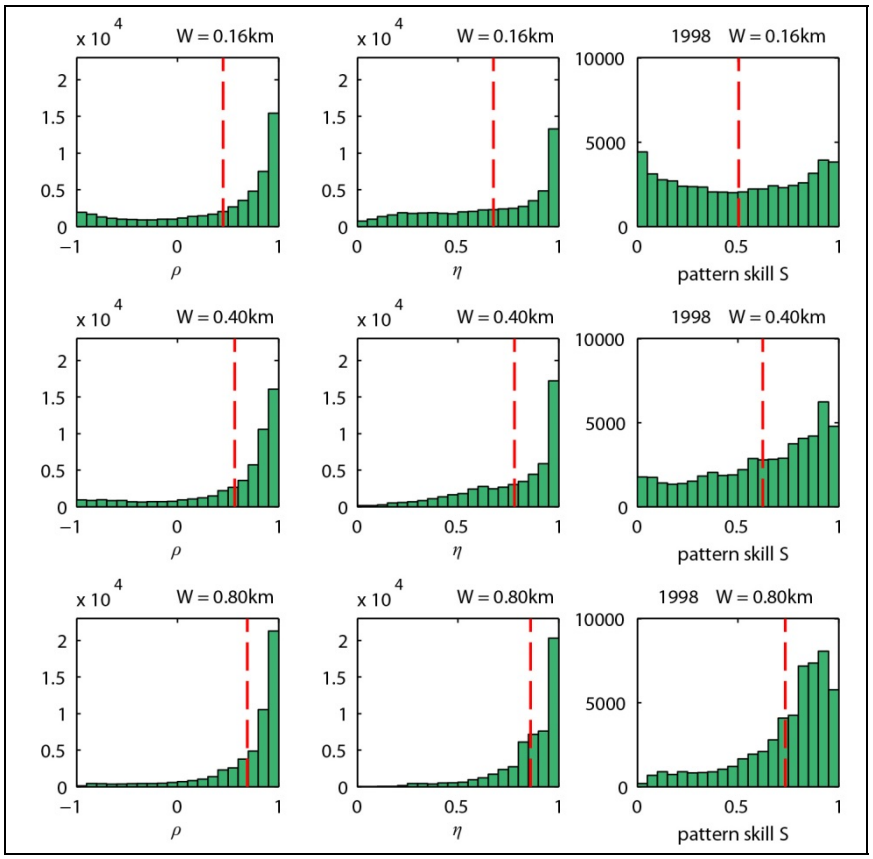

Figure 6. Histograms of the correlation, amplitude similarity and pattern skill for the three window sizes *W* = 0.16, 0.4 and 0.8km. Note that the histograms correspond to the respective maps in Fig. 5. The red lines indicate the domain-averaged values which can be seen to increase with spatial scale.

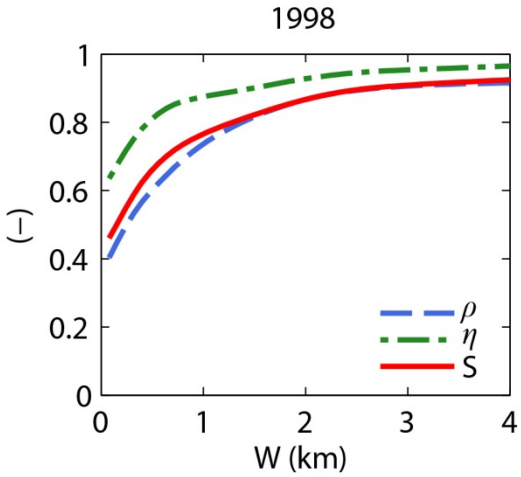

Figure 7. Structural and amplitude similarity and pattern skill as a function of window size.

Information content versus skill

Output of high-resolution morphodynamic area models is generally presented at the resolution of the computational grid. The previous findings suggest, however, that the high-resolution detail may not be skilful. Consequently, a smoother bathymetry (Fig. 8) may be more skilful than the original, computed bathymetry (Fig. 4). The bathymetries in the left and right columns of Fig. 8 are obtained by applying a moving average to the original bathymetries, using window sizes of W = 0.4 and 1.6km, respectively (using Eqs. 3-4).

To determine the effect of leaving the high-resolution detail out, we apply the same validation procedure as before, at a range of window sizes, but now not to the full-resolution bathymetries, but to their smoothed counterparts. The aggregated results are shown in Fig. 9. For clarity, the skill trend for the full-resolution bathymetries (Fig. 7) is repeated in Fig. 9. The latter figure confirms that for all scales the presented smoother bathymetries are more skilful. Note that for the bathymetries smoothed with W = 0.4km, all scales have a skill around or above the target skill of 0.7.

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

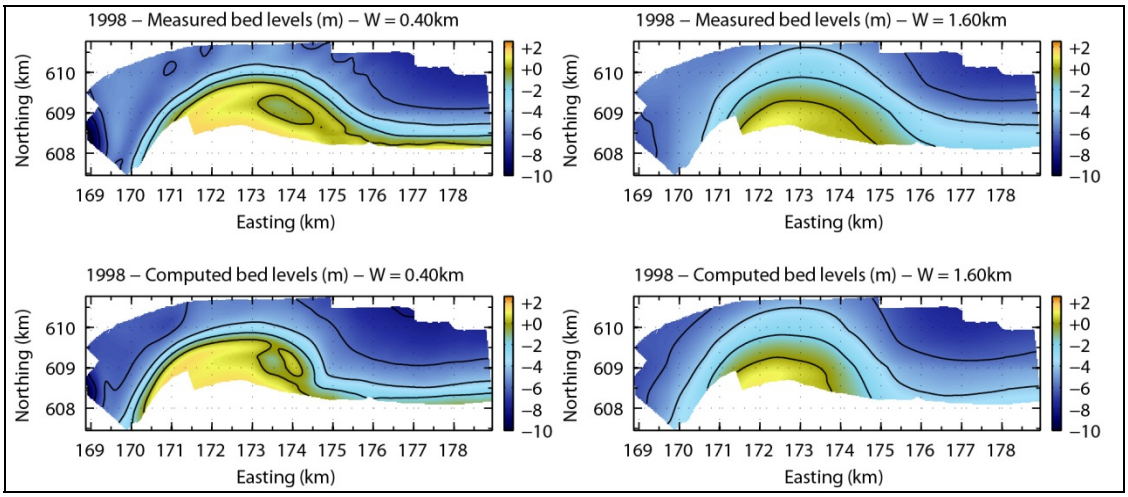

Figure 8. Spatial means, obtained by Eqs. 3-4, of the original high-resolution bathymetries. Left: W = 0.4km and right: W =1.6km.

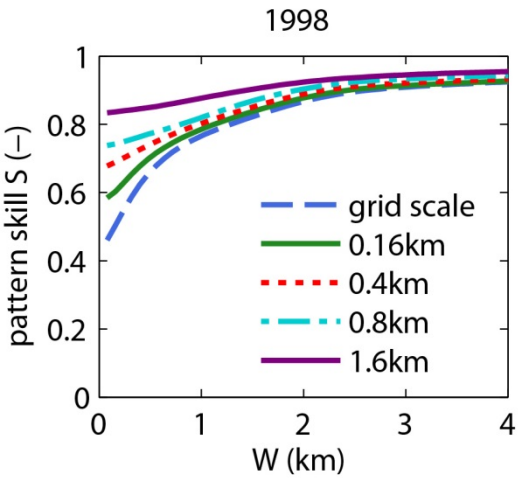

Figure 9. Pattern skill versus window size for bathymetries with a different level of smoothening (a moving average at window sizes ranging from 0.16km to 1.6km. The pattern skill at full resolution (Fig. 7) is repeated here and indicated with 'grid scale'.

CONCLUSIONS

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

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

In summary, the method can be used to:

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

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

REFERENCES

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

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

Bosboom, J., A.J.H.M. Reniers, and A.P. Luijendijk. In press. On the perception of morphodynamic model skill, accepted for publication in *Coastal Engineering*. http://dx.doi.org/10.1016/j.coastaleng.2014.08.008.

Fotheringham A.S., C. Brunsdon C., and M. Charlton. 2002. Geographically weighted regression: the analysis of spatially varying relationships, Wiley, New York, 284 pp.

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

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

Roberts, N.M and H.W. Lean. 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136.1, 78-97. http://dx.doi.org/10.1175/2007MWR2123.1.

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

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

Wang, Z., E.P. Simoncelli, and A.C. Bovik. 2004. Multiscale structural similarity for image quality assessment, *Proceedings of the 37th Conference on Signals, Systems and Computers*, 2003, 2, 1398-1402. http://dx.doi.org/10.1109/ACSSC.2003.1292216.