Research Article

Are Feature Agreement Statistics Alone Sufficient to Validate Modelled Flood Extent Quality? A Study on Three Swedish Rivers Using Different Digital Elevation Model Resolutions

Nancy Joy Lim and Sven Anders Brandt

Department of Computer and Geospatial Sciences, University of Gävle, SE-801 76 Gävle, Sweden

Correspondence should be addressed to Nancy Joy Lim; naylim@hig.se

Received 26 April 2019; Revised 22 August 2019; Accepted 13 September 2019; Published 13 November 2019

Guest Editor: Raffaele Albano

Copyright © 2019 Nancy Joy Lim and Sven Anders Brandt. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Hydraulic modelling is now, at increasing rates, used all over the world to provide flood risk maps for spatial planning, flood insurance, etc. This puts heavy pressure on the modellers and analysts to not only produce the maps but also information on the accuracy and uncertainty of these maps. A common means to deliver this is through performance measures or feature statistics. These look at the global agreement between the modelled flood area and the reference flood that is used. Previous studies have shown that the feature agreement statistics do not differ much between models that have been based on digital elevation models (DEMs) of different resolutions, which is somewhat surprising since most researchers agree that high-resolution DEMs are to be preferred over poor resolution DEMs. Hence, the aim of this study was to look into how and under which conditions the different feature agreement statistics differ, in order to see when the full potential of high-resolution DEMs can be utilised. The results show that although poor resolution DEMs might produce high feature agreement scores (around $F > 0.80$), they may fail to provide good flood extent estimations locally, particularly when the terrain is flat. Therefore, when high-resolution DEMs (1 to 5 m) are used, it is important to carefully calibrate the models by the use of the roughness parameter. Furthermore, to get better estimates on the accuracy of the models, other performance measures such as distance disparities should be considered.

1. Introduction

Flood risk maps have become standard documents in local, regional, national, and sometimes even international spatial planning. They are widely used and considered as valuable sources of information, mainly in planning and development projects. Detailed positioning of houses, buildings, and infrastructure is decided based on these maps. However, the reliability of these flood maps has been questioned for years. Deterministic flood maps, in particular, which show defined flooded areas associated with a specific flood-return year (e.g., the 100-year flood), are criticised to mislead map users to think that they show absolute boundaries of flood extent. Nonetheless, they can never be exact due to effects of different factors and modelling assumptions (e.g., input data, model, and model parameters) in the production of these maps. Different data, models, or parameters can change the position of the flood borders.

To overcome this limitation of flood model results, they are calibrated (e.g., using different roughness values) against a historic (reference) flood event with the same discharge. Each flood extent from the calibration is then validated for its accuracy through a performance (goodness-of-fit) measure that quantifies the model quality. Numeric scores provided by the performance measures are used to justify the choice of an optimal model that can be used as a flood hazard map. In uncertainty and probabilistic mapping implemented through the Generalised Likelihood Uncertainty Estimation (GLUE) methodology [1], these performance estimators play an important role, where they are used for deriving the
likelihood weights assigned to the results of the ensemble, prior to the creation of the map [2]. Moreover, performance scores are widely used to analyse the sensitivity of the modelled flood to the effects of the input topographic data [3–6], roughness value [7–11], and hydraulic models, or their numerical solutions [3, 7, 8].

In flood extent validation studies, performance measures used include various types of feature agreement (F) statistics [9], as well as disparity measures [2]. F-statistics are the most commonly applied in assessing flood model predictions. They give a general view of the model’s performance by considering the total sizes of over- and underestimation and overlap between the model and the reference data [12], but the outcome may be affected by what is prioritised in the equation. The disparity measures (mean and median disparities), on the contrary, account for the distance between the modelled and reference data at each point sample location [2, 12]. Nevertheless, the scores will be dependent on where the samples are taken or the sampling strategy employed. According to Beven [13], all these equations have their assumptions, which can have implications on the choice of optimal models. Hence, decisions of the optimal model can be relative to the performance equation used.

Lim and Brandt [12] also showed that the combination of the digital elevation model (DEM) and roughness (Manning’s n) can influence the performance measures such as the F-statistics scores. However, it remains uncertain how these various statistical estimators’ results differ and when they differ. Hence, the aims of this study are to evaluate the following: (1) how the different performance measures, such as the different feature agreement statistics, vary in their results, in terms of the quantified values they produce; (2) how the generated flood extents of models considered as “best” coincide with the validation data and the computed performance scores; (3) how the sizes of overlap and under- and overestimation vary with the given DEM or DEM and roughness pairs; and (4) how the proportion of areas correctly and incorrectly modelled agree with the performance measures.

The effects of model performance in the assessment of simulation results based on the influence of the DEM alone, or both the DEM and roughness values used, are investigated in relation to the aims. The topographic data remains an important hydraulic numerical model input that provides the geometry of the study area (through, e.g., cells, triangular meshes or cross sections) and is the main source of the elevation values [2]. The grid size or the resolution mainly affects the details in the modelling and the accuracy (both vertical and horizontal) of model results [4, 5]. Nonetheless, as what is shown in some studies [3, 6, 12], even lower resolution DEMs (≥25 m) can produce better quantified performance. Thus, basing decisions on optimal models from a given performance measure alone may discard the benefits of using fine resolution DEMs. In addition, as the Manning’s n is the parameter that is adjusted during the calibration, it is important to also see how these values, especially when paired with different DEMs, can affect the performance scores.

2. Verification Measures for Spatial Predictions of Floods

One way to distinguish the goodness of predictions from simulation result or modelling is through its quality or agreement to the observed data (truth) [14]. This quality is assessed through verification methods that help describe or quantify their performance. A model is usually compared with observational data, based on binary conditions (i.e., yes/presence or no/absence), through a 2 × 2 contingency table [15]. In Table 1, hits (A) represent predicted events to happen and did take place. False alarms (B) are events modelled to occur but did not occur, whereas misses (C) are events that actually happened but were not modelled to happen. Model results that are correctly predicted not to happen fall under correct rejection (X). A number of performance estimators based on this table are described in Mason [15].

In flood inundation studies, the goodness of model prediction can be evaluated by assessing the overall, so called global, spatial accuracy of the simulated flood result produced from model calibrations or ensemble modelling, against a reference dataset, which is often a historical flood extent having the same magnitude as the modelled one. In such applications where the quality of spatial predictions of modelled events are verified, performance measures that can be applied will include accuracy, the critical success index (CSI), hit rate or the probability of detection (POD), false alarm rate or the probability of false detection (POFD), and the false alarm ratio (FAR) [9].

The degree of correspondence between the model and the observation data can be estimated through computing the proportion correctly predicted by the model (i.e., accuracy) [9, 15]:

\[
\text{accuracy} = \frac{A + X}{A + B + C + X}
\] (1)

According to Schaefer [16], this equation can be affected by the size or number of correct rejections, especially if it is large, which can lead to high accuracy score. In flood modelling, where dry areas are usually larger, a high accuracy score can be achieved, despite the low number of hits in the model [10]. Thus, in most performance measures used, correct rejection (X) is disregarded. The removal of the correct rejection from the original accuracy equation leads to another performance measure widely known as the critical success index or the threat score:

\[
\text{CSI} = \frac{A}{A + B + C}
\] (2)

In flood model verification, the application of CSI allows to focus on the correct prediction made by the model (hits, A), and the two main causes of flood modelling errors, i.e., false alarm (B) and misses (C). There are several equations based on the CSI, and they are known as the feature agreement (F) statistics. Generally, they are implemented by either considering (1) the binary flooding conditions (i.e., flooded = 1 or dry = 0) of each cell (in a raster data) in the observed (D) and modelled (M) result [9, 17] or (2) by comparing the areal size and location of the modelled flood
against the reference data [18] (in vector data). All \( F \)-statistics equations rely on three important factors in the contingency table that help assess how well the model forecast the flooding. These are (1) the total number of cells \((P)\) correctly modelled to be flooded \((M_1D_1)\) or the size of overlap \((A, \text{hits})\) between the modelled and the reference data, (2) total number of cells \((M_1D_0)\) or size \((B)\) overestimated \((\text{false alarms})\) by the model, and (3) total cells \((M_0D_0)\) or size underpredicted \((C, \text{misses})\).

The most typically used \( F \)-statistics is \( F1 \), which is based on the original CSI equation. It is based on getting the intersection \((\text{hits})\) and union \((\text{false alarm} + \text{misses})\) of the modelled and observed event [3] (equation (3)). The values generated by the equation is from 0 (no fit) to 1 (perfect model fit):

\[
F_1 = \frac{M \cap D}{M \cup D}
\]  

(3)

However, this feature agreement statistics is also known to give bias in the results it produces [9], especially when the prediction gives large overlap size \((\text{hits})\) [12]. If the flooded area is large, as is the case when rivers flood wide flat floodplains, the modelled flood area will overlap the reference flood area to a relatively big extent, compared with the areas that are either over- or underestimated. This means that a rather poor model can achieve strong statistics. As stated by Lim and Brandt [12], as long as the model has \( \text{hits} \) greater than the combined sizes of \( \text{misses} \) and \( \text{false alarms} \), the equation will lead to an \( F \)-value > 0.50. Because of this limitation, several variants of the equation have been introduced (Table 2) to improve the quantified scores. These versions of the equation modify the numerator to penalise the model with the overprediction \((F_{2o})\), underprediction \((F_{2u})\), or both under- and overpredictions \((F_3)\) it produces [9]. Among these variants, \( F_{2o} \) is a widely used performance estimator in flood extent validation studies and in creating probabilistic maps. According to Hunter [9] and Hunter et al. [8], it constrains the overlap produced by the model in \( F_1 \), by penalising it with the overprediction it produces. \( F_{2u} \) and \( F_3 \) are the least applied among the four equations. In \( F_{2u} \), the overlap is subtracted by the underestimation made by the model, whereas in \( F_3 \), both under- and overestimation are subtracted from the overlap. Maximum value of these performance measures is also 1, indicating a perfect model fit with the reference data. Their minimum can be \(-1\).

Other spatial performance measures that can be used in flood modelling are the \textit{hit rate} or \textit{probability of detection}, \textit{probability of false detection}, and \textit{false alarm ratio}. The \textit{POD} is a standard statistics that indicates the ratio of correct prediction by the model, to the total observed data [15, 16]. In flood model verification, it tells the total number of pixels or areal size correctly predicted by the model to be flooded, in relation to the reference data, by considering both the overlap and the underestimation made by the model. Maximum \textit{POD} score is 1 in models with perfect overlap with the reference data. According to Mason [15] and Hunter [9], the \textit{POD} measure alone cannot be used as a performance estimator because the equation is affected by the size of the \textit{hits} \((A)\). Large hits (and low underestimation) can generate high performance score, even if the model has largely overestimated the flooding. Thus, the \textit{POD} equation should be supplemented with information that considers errors in overestimation [9], through the \textit{probability of false detection} (also known as the \textit{false alarm rate}) and the \textit{false alarm ratio}. \textit{POFD} accounts for the proportion of overprediction made by the model in relation to the \textit{correct rejections} \((X)\), while the \textit{FAR} score considers the proportion of overestimation in relation to the \textit{hits}. Maximum scores for both performance measures are 0, i.e., no overestimation made by the model. However, these two can also pose similar problems as the \textit{POD}. As the \textit{POFD} is related only to \textit{correct rejections}, how well the inundation is produced in terms of the overlap or how the underprediction is minimised are not considered in the equation. On the contrary, although \textit{FAR} accounts for the effect of the size of overlap made by the model, the size underpredicted is not accounted in the equation.

A further measure that can be applied to spatial model verification is the \textit{BIAS}. This measure is for quantifying the consistency between the model result and the observed flood event [9] but does not indicate the model skill or performance [15]. Therefore, the bias score is considered more of a descriptive statistics that specifies the ratio of the predicted data to the observation [15]. In flooding applications, this measure indicates if overall the model result underestimates \((\text{BIAS} < 1)\) or overestimates \((\text{BIAS} > 1)\) the flood. A value of 1 shows no bias. The farther away the score from 1, the larger the bias on either under- or overestimation.

### 3. Materials and Methods

Model performance measures are evaluated for model results based on different DEM resolutions in combination with different roughness values or single-handedly by the DEM resolution. To meet the aims, there were three study areas at different geographical locations in Sweden used (Figure 1): Voxna, Testebo, and Eskilstuna (two different parts) (Figure 1). Voxna and Testebo rivers were used to determine the influence of DEM together with Manning’s \( n \) on the performance scores. In analysing the effect of the DEM resolution (including its errors), the Eskilstuna river was used as a case. Simulation results from earlier studies of Lim and Brandt [12] (Voxna and Testebo), Klang and Klang [19], and Brandt [20, 21] (Eskilstuna) were used in the current study. Specific data processing and modelling procedures are thereby described in the said articles in detail.
Table 2: Different flood verification measures implemented for raster and vector data (revised from Hunter [9]).

| Verification measures | Binary classification (raster-based) | Areal size (vector-based) | Values |
|-----------------------|--------------------------------------|---------------------------|--------|
| **Feature agreement statistics:** | | | |
| $F_1$ | $\frac{\sum P_{D1} - \sum P_{D0} \sum P_{D1}}{\sum P_{D1} + \sum P_{D0} \sum P_{D1}}$ | $\frac{A}{A+B+C}$ | 0 to 1 |
| $F_{2O}$ | $\frac{\sum P_{D1} - \sum P_{D0} \sum P_{D1}}{\sum P_{D1} + \sum P_{D0} \sum P_{D1}}$ | $\frac{A-B}{A+B+C}$ | $-1$ to 1 |
| $F_{2U}$ | $\frac{\sum P_{D1} - \sum P_{D0} \sum P_{D1}}{\sum P_{D1} + \sum P_{D0} \sum P_{D1}}$ | $\frac{A-C}{A+B+C}$ | $-1$ to 1 |
| $F_3$ | $\frac{\sum P_{D1} - \sum P_{D0} \sum P_{D1}}{\sum P_{D1} + \sum P_{D0} \sum P_{D1}}$ | $\frac{A-B-C}{A+B+C}$ | $-1$ to 1 |
| **Hit rate/probability of detection (POD)** | | | |
| | $\frac{\sum P_{D1}}{\sum P_{D1} + \sum P_{D0} \sum P_{D1}}$ | $\frac{A}{A+B+C}$ | 0 to 1 |
| **Probability of false detection (POFD)/false alarm rate** | | | |
| | $\frac{\sum P_{D0}}{\sum P_{D1} + \sum P_{D0} \sum P_{D1}}$ | $\frac{B}{A+B+C}$ | 0 to 1 |
| **False alarm ratio (FAR)** | | | |
| | $\frac{\sum P_{D0}}{\sum P_{D1} + \sum P_{D0} \sum P_{D1}}$ | $\frac{B}{A+B+C}$ | 0 to 1 |
| **BIAS** | | | |
| | $\frac{\sum P_{D0} \sum P_{D1}}{\sum P_{D1} + \sum P_{D0} \sum P_{D1}}$ | $\frac{A+B}{A+B+C}$ | 0 to co |

Figure 1: Locations of the study areas ©Lantmäteriet.
3.1. Study Areas, Data, and Hydraulic Modelling

3.1.1. Voxna River. The first site is the Voxna river, situated in Edsbyn. The total size of the study area is about 14.7 km². Its western part is flat, consisting mainly of agricultural lands and forests. At the east, the river flows through the town of Edsbyn, where mostly built-up areas are present. This part of the river is steeper. The Voxna river has a mean annual flow of 40 m³/s at the outlet.

Topographic data used for creating the digital elevation models for Voxna came from a combination of bathymetric and LiDAR data. The DEMs were produced from the topographic data through TIN generation and conversion to raster, which is the main input required by the hydraulic model used [12]. The different DEMs used for this river had resolutions of 3, 4, 5, 10, 15, 20, 25, and 50 m. The highest resolution DEMs were not possible to be used for the study area due to the technical limitations of the 2D hydraulic model in terms of the size of the site. For each DEM, 10 different Manning’s roughnesses were used, ranging from 0.010 to 0.100, with 0.01 increments in accordance with Horritt and Bates [7]. This was mainly to see how these different roughness coefficients affect the model results from different DEM resolution models.

The simulated flow was 360 m³/s, which is equal to the big flood event in 1985. A corresponding reference flood inundation map was provided by Ovanåker municipality for this particular event. The flow was modelled as steady state 2D flow in CAESAR-LISFLOOD software (Coulthard et al. [22]). Eighty simulations from the different combinations of DEM and Manning’s n were carried out for the river.

3.1.2. Testebo River. The second area is around the Testebo river at Forsby, north of Gävle. It has a mean annual flow of 12 m³/s. The surrounding areas of the river, particularly its central portion is mainly flat arable land, with some mixed forests. Most residential areas are situated at the edges of these lands. Steeper sections of the river are located at the eastern and southern parts of the study site.

DEM produced for Testebo river were processed in the same way as for the Voxna river. Similar DEM resolutions (except for the 1 and 2 m resolution, which were also included) and range of roughness values as for the Voxna river were used for the modelling. Here, the simulated flow was 160 m³/s, which coincides with the 100-year flood. Gävle municipality provided a flood inundation map over the 1977 flood, which is equivalent to this flow. A steady state 2D flow using the CAESAR-LISFLOOD software was also implemented to the study area [12]. A total of 100 simulations were conducted for the combinations of resolution and roughness values.

3.1.3. Eskilstuna River. The third location consists of two parts of the Eskilstuna river (with a mean annual flow of 24 m³/s) in Eskilstuna. The northern part of the study area is located in the city centre where most residential, commercial, and industrial establishments are concentrated. This site is characterised by relatively steeper banks. The other location is situated just south of the city, where the floodplain is flatter. Residences are visible in the western part of this study site, along the river. Some forests are also lining up the river. At the edges of these forested areas are arable lands.

DEM resolutions used for the modelling were 1.04, 2.09, 3.13, 10, 25, and 50 m. Note, however, that these particular DEMs have not been derived in the same way as those for Voxna and Testebo rivers, where linear interpolation of point cloud data was used to construct the DEMs of the desired resolution. Here, the original point cloud data, which had a density of 1.64 and 1.36 points/m², corresponding to cell sizes of 0.78 and 0.86 m for the flat and steep subareas, respectively, were used to produce the reference elevation model. This DEM was then resampled and degraded by introducing random errors to simulate the behaviour of the results from a Leica ALS50 LiDAR instrument for different flight heights. Details of the DEM degeneration can be found in Klang and Klang [19] and Brandt [20, 21].

A single roughness value (0.033) was used for all model runs. This Manning’s n is the most commonly suggested friction coefficient for rivers in Sweden. As a single Manning’s n was used, this means that the uncertainties of modelled flood extents can be isolated to only depend on the characteristics of the DEM and not on the DEM resolution in combination with roughness [20, 21]. Flow equivalent to the maximum probable flow of 198 m³/s was used for the simulations. A steady state 1D flow was simulated using the HEC-RAS software (Hydrologic Engineering Center [23]).

3.1.4. Model Performance Evaluation. Different feature agreement statistics, which are applied in flood extent validation studies, were used to assess the overall performance of the simulation results. Each model result was compared to the observed flood data (i.e., the historical flood maps for Voxna and Testebo rivers and the simulation result from the original point cloud data in case of Eskilstuna). Since the reference data were all in vector formats, equations accounting for areal size (cf. Table 2) were used. This was implemented by overlaying the reference and the given modelled flood map resulting from simulating a given DEM, or DEM and Manning’s n combination. Then, the sizes of overlap and under- and overestimation were taken (Figure 2), prior to the application of the specific F-statistics equation.

Aside from using the F-statistics, each simulation and best model results from the different resolution DEMs were also evaluated using two other spatial performance measures that can help understand the performance scores provided by the F-statistics, in terms of the probability of detection and false alarm ratio. Both statistics were considered to be able to determine how well the model produces the inundation in relation to the hits, as well as underestimation it produces (POD) [9], and how it overpredicted the results (FAR). Probability of false alarm or the false alarm rate was not used in this study, due to its inclusion of correct rejection (X) in the equation, which is difficult to account in spatial model results such as in flood extents because of huge dry areas being modelled [10]. To further understand whether the
model from a given performance, DEM, and DEM and Manning’s n combination tends to over- or underestimate the flood predictions, BIAS scores were also computed.

4. Results

The effects of performance in evaluating quality of flood models’ results as influenced by both the DEM and Manning’s n used (Testebo and Voxna), and by the DEM (Eskilstuna, North and South), were analysed through the following: (1) the overall performance scores of the simulation results, (2) the most optimal models’ performance scores and their corresponding spatial extents and size of inundation areas, (3) changes in the sizes of overlap and over- and underestimation produced by the model for the given DEM and DEM and roughness combinations, and (4) the POD, FAR, and BIAS scores of the models.

4.1. Combined Effects of DEM and Roughness Values (Testebo and Voxna Rivers)

4.1.1. Overall Performance Results. For both Testebo and Voxna rivers, various DEM resolution together with different roughness values were evaluated to find out how they affect the quantified performance scores using the different feature agreement statistics. From the 3D surface plots in Figure 3, it can be seen that the DEMs produced varying performance values in each measure, depending on the Manning’s n paired with them. Moreover, from the two study sites, the Testebo river produced more variation in the performance score (for each goodness-of-fit measure) when using high-to-low-resolution DEMs, compared with the Voxna river. Overall, the Voxna river resulted in higher performances from the different combinations of DEMs and roughnesses, with only 1 to 4 model results (out of 80) having $F < 0.50$ in the four $F$-statistics. It can also be observed that, regardless of the goodness-of-fit measure used for both study areas, higher performances were more concentrated in medium to lower resolution DEMs (from 10 to 50 m), paired with a wider range of Manning’s n values, whereas better results from higher resolution DEMs (1–5 m) came from a smaller range of roughness values.

Among the different feature agreement statistics, $F2_O$ had fewer simulation results that gave higher performance. Lower roughness values also led to better results for this measure, especially when paired with coarser resolution DEMs. While $F2_U$, more simulation results produced higher performances. Nevertheless, it can also be seen that higher resolution DEMs performed poorly with lower Manning’s values (particularly for the Testebo river, producing more negative performances), while coarser resolution DEMs performed well with most of Manning’s n, especially for the Voxna river. The performance patterns for F1 and F3 were the same, but the latter produced lower and even negative values, especially for the higher resolution DEMs paired with lower roughness.

4.1.2. Most Optimal Performance Results. The four performance measures considered different pairs of DEM and Manning’s n to be the most optimal from the ensemble of 80 (Voxna) and 100 (Testebo) simulations (Figure 4). In general, $F2_O$ produced the best results from each DEM with much lower roughness values, while $F2_U$ produced the best results with higher friction. F1 and F3, which had the same optimal model results, almost followed the roughness pattern of $F2_U$ for each resolution, but using a lower friction coefficient. It is also visible from the diagram that the best models from the three performance estimators (F1, $F2_U$, and F3) that used higher resolution DEMs (1–5 m) worked well with higher roughness, and that the most optimal Manning’s n used decreased as the resolution became coarser.

When considering the maximum performance scores for the different DEMs, the lowest resolution DEMs (25 and 50 m) produced the most optimal performance in all four $F$-statistics for the Testebo river ($F1_{50 m,n=0.030} = 0.730$, $F2_{50 m,n=0.040} = 0.651$, $F2_{50 m,n=0.040} = 0.646$, and $F3_{50 m,n=0.040} = 0.460$). For the Voxna river, the 10 m DEM got the highest scores when using $F1_{n=0.030} = 0.850$, $F2_{U,n=0.030} = 0.807$, and $F3_{n=0.040} = 0.70$; while the 5 m DEM received the highest with $F2_{O,n=0.030} = 0.763$. Although the maximum performance range did not vary much within each $F$-statistics (Tables 3 and 4), the variations in extents using the different DEMs and Manning’s roughness were more discernible when compared with each other visually (Figure 5). Flood extents having the highest $F2_U$ values produced big underestimations particularly for the highest resolution DEMs in both rivers. The 1 to 5 m DEMs underestimated the southern part of the Testebo river and the southeastern section of Voxna river. The 10 to 50 m DEMs underpredicted the flooding in the north of Testebo, while in Voxna, these resolutions overpredicted the
flooding, $F_2u$, produced larger extents in all DEM resolutions for Testebo (especially in the south and southeast) than $F_1|F_3$ (except for the 50 m DEM) because of the higher roughness values used. In Voxna, the difference in extents between $F_2u$ and $F_1|F_3$ were less noticeable in the map, but if the size of the inundation area will be compared, the former had a larger flooded area as brought by using the higher roughness value (i.e., $n = 0.030$) (Table 4).

Figure 3: Different performance results for the Testebo and Voxna rivers using various combinations of DEM resolution and roughness parameter.
fl_hesizeoffloodedareasproducedbythebestmodelsforeachresolutionandperformancemeasureandthepercentdifferenceinarealsizeinrelativetothereferencedataarefurthershowninTables3and4.Apercentageequalto0meansthatthesizeoftheinundationareafromthesimulationisthesameastherefERENCEdata.Takenotethatthisvaluedoesnotsignifyanexactmatchinneighbors(overlap)betweentheModelandthereference.Anegativevalue(>0)meansthatthemodeledfloodareaismallertexthanthereference data,whileifitispositive(>0),themodeledsizeislarger.Thefartherawaythevalueisfrom0meansthe smallerorlargertheunder-orextrationis.Here,theclearlyshowthatmodelscomputedwiththehighest

Table 3: Performance scores of most optimal DEM resolutions (with associated roughness in small font) for the Testebo river, together with inundation areas and the percent differences in size from the reference data.

| Res. (m) | F1 and F3 | Flood area (km²) | % diff. | Max. score | Flood area (km²) | % diff. | Max. score | Flood area (km²) | % diff. |
|---------|-----------|------------------|---------|------------|------------------|---------|------------|------------------|---------|
| 1       | 0.6680.080| 0.3370.080       | 0.674   | 17.42      | 0.4680.050       | 0.41    | −27.00     | 0.6070.100       | 0.789   |
| 2       | 0.6690.080| 0.3390.080       | 0.683   | 18.99      | 0.4670.050       | 0.428   | −25.44     | 0.6010.090       | 0.743   |
| 3       | 0.6700.070| 0.3400.070       | 0.691   | 20.38      | 0.4650.070       | 0.624   | 8.71       | 0.6040.090       | 0.752   |
| 4       | 0.6710.070| 0.3410.070       | 0.635   | 10.63      | 0.4940.040       | 0.311   | −45.82     | 0.6080.090       | 0.760   |
| 5       | 0.6740.070| 0.3450.070       | 0.639   | 11.32      | 0.4720.040       | 0.29    | −49.48     | 0.6050.090       | 0.762   |
| 10      | 0.6940.070| 0.3880.070       | 0.512   | −10.80     | 0.5900.070       | 0.512   | −10.80     | 0.6220.080       | 0.733   |
| 15      | 0.6970.060| 0.3950.060       | 0.643   | 12.02      | 0.5250.040       | 0.367   | −36.06     | 0.6380.080       | 0.782   |
| 20      | 0.7150.060| 0.4300.060       | 0.575   | 0.17       | 0.6160.050       | 0.476   | −17.07     | 0.6300.080       | 0.742   |
| 25      | 0.7210.040| 0.4420.040       | 0.487   | −15.16     | 0.6510.040       | 0.487   | −15.16     | 0.6450.070       | 0.790   |
| 50      | 0.7300.030| 0.4600.030       | 0.573   | −0.17      | 0.5960.030       | 0.573   | −0.17      | 0.6460.010       | 0.574   |

The size of flooded areas produced by the best models for each resolution and performance measure and the percent difference in area size in relation to the reference data are further shown in Tables 3 and 4. A percentage equal to 0 means that the size of the inundation area from the simulation is the same as the reference data. Take note that this value does not signify an exact match in extents (overlap) between the model and the reference. A negative value (>0) means that the modelled flood area is smaller than the reference data, while if it is positive (>0), the modelled size is larger. The farther away the value is from 0 means the smaller or larger the under- or overestimation is. Here, the values clearly show that models computed with the highest

Table 4: Performance scores of most optimal DEM resolutions (with associated roughness in small font) for the Voxna river, together with inundation areas and the percent differences in size from the reference data.

| Res. (m) | F1 and F3 | Flood area (km²) | % diff. | Max. score | Flood area (km²) | % diff. | Max. score | Flood area (km²) | % diff. |
|---------|-----------|------------------|---------|------------|------------------|---------|------------|------------------|---------|
| 3       | 0.8230.080| 0.6460.080       | 5.410   | 8.80       | 0.7390.030       | 4.331   | −12.91     | 0.7870.090       | 5.585   |
| 4       | 0.8230.080| 0.6470.080       | 5.457   | 9.75       | 0.7340.040       | 4.479   | −9.92      | 0.7870.090       | 5.605   |
| 5       | 0.8430.080| 0.6870.080       | 5.303   | 6.65       | 0.7630.050       | 4.410   | −11.31     | 0.8020.080       | 5.303   |
| 10      | 0.8500.040| 0.7000.040       | 5.241   | 5.40       | 0.7570.030       | 5.086   | 2.29       | 0.8070.040       | 5.241   |
| 15      | 0.8410.040| 0.6820.040       | 5.263   | 5.84       | 0.7380.020       | 5.051   | 1.59       | 0.8050.060       | 5.550   |
| 20      | 0.8430.030| 0.6850.030       | 5.176   | 4.10       | 0.7440.020       | 5.040   | 1.35       | 0.8030.040       | 5.339   |
| 25      | 0.8230.020| 0.6570.020       | 5.300   | 6.59       | 0.7150.010       | 5.154   | 3.65       | 0.7900.030       | 5.438   |
| 50      | 0.8280.010| 0.6570.010       | 5.407   | 8.73       | 0.6970.000       | 5.407   | 8.73       | 0.7940.030       | 5.556   |

F2_U scores have bigger modelled area size than the reference data, in both study sites, and the other three performance measures (except for the 50 m DEM for the Testebo river). Inundation areas of best models derived from this performance measure for the Testebo river have size differences from the reference data ranging from 27 to 37%. For Voxna, the differences in sizes were smaller (5−12%). In the Testebo river, the 50 m DEM that was paired with n = 0.010 produced a size equal to the reference data (0% difference), but it can be seen in Figure 3 that they did not exactly match each other’s extents. Results with the highest F2_U have generally smaller inundation areas, especially for the Testebo river. But again, for this river, although the area size of inundation
produced from the 50 m DEM was smaller than the actual flood data, the percent difference was very small (-0.17%) compared with the other DEMs. In addition, the most optimal results based on the highest resolution DEMs for this performance measure have smaller inundated areas than the reference (-25 to -49% difference for the 1, 2, 4, and 5 m for Testebo and -9 to -13% for Voxna). With $F_1$ and $F_3$, all optimal models produced larger flood extents than the reference data in Voxna, whereas in Testebo, the 10, 25, and 50 m DEMs produced smaller flood areas than the rest of the DEMs.

![Figure 5: Extent comparison of most optimal results produced for each performance measure. $F_1$ and $F_3$ have the same results although the scores were lower for the latter.](image)
4.2. Resolution’s Effect on Quantified Performance (Eskilstuna River). For the Eskilstuna river, a constant roughness ($n = 0.033$) was used for all DEMs to produce the results. Irrespective of the topographic characteristics of the study area, i.e., steep (north) or flat (south), the highest resolution data (1 m) received the highest performance in all four measures (Figure 6 and Table 5). A distinct trend is also shown that the performance decreased as the resolution became coarser. Among the goodness-of-fit measures, $F_1$ showed a more gradual decrease in performance from finer to coarser resolution, while $F_3$ manifested a more significant decrease. From the two study areas, it can be noticed that the shift in performance between the highest resolution and the 10 m DEM was more prominent for the northern part of the river (i.e., from 0.93–0.96 down to 0.74–0.87), particularly when $F_{2_U}$ and $F_3$ were used.

When looking at the inundation areas produced by the different resolution DEMs (Table 5), 50 m produced the largest flood areas in both sites although there was larger difference in the north (10.55%). Also noticeable is that 3 m data produced the next largest difference after 50 m. Lowest difference in the north (10.55%). Also noticeable is that 3 m was at 10 m DEM was more prominent for the northern part of the river (i.e., from 0.93–0.96 down to 0.74–0.87), particularly when $F_{2_U}$ and $F_3$ were used.

When looking at the inundation areas produced by the different resolution DEMs (Table 5), 50 m produced the largest flood areas in both sites although there was larger difference in the north (10.55%). Also noticeable is that 3 m data produced the next largest difference after 50 m. Lowest percent difference in areal size was made by 1 m DEM, both in the northern and southern parts of the river. The 25 m also generated smaller percent difference in flooded area than some of the higher resolution DEMs.

When the inundation extents are examined (Figure 7), there were higher degree of overlaps (with the reference) produced by the higher resolution (1–10 m) DEMs. But after these resolutions, the mismatches at the edges became more recognisable, especially when 50 m was used. Over- and underestimations in flooding in some parts of the study areas also became obvious with this resolution.

4.3. Changes in Sizes of Overlap and Over- and Underestimation for the Different Resolution DEMs. All performance measures’ equations are based on accounting for the overlap, overestimation, and underestimation sizes produced by the simulation results. If these three will be looked at for each of the results from the different study cases (Figure 8), and for each DEM resolution, it can be seen that the general pattern for the Voxna and Testebo rivers were similar; i.e., there were more overlaps attained when using the lower resolution data and high roughness. Nonetheless, these resolutions also generated more overpredictions (but minimal underpredictions), especially when they were paired with higher roughness values. On the contrary, the sizes of overlaps produced by higher resolution DEMs were generally lower than the coarser resolution DEMs. Furthermore, higher friction coefficients paired with them generated better overlap size, as well as overestimations, while lower frictions produced underestimations. The big shifts in the size of overlap and over- and underestimation from the highest (1 m) to the lowest (50 m) DEMs were most obvious for the Testebo river.

In both the southern and northern parts of Eskilstuna river, the overlap sizes did not vary much among the results from the different digital elevation models, but 1 m got the highest inundation match with the reference data. In addition, the overlap size decreased as resolution became coarser. The increase in size of overestimation from finest to coarsest DEM was similar to the Testebo and Voxna rivers. Nevertheless, the pattern that followed for the underestimation was opposite the two other study areas; i.e., the coarser resolution DEMs generated more underpredictions for the Eskilstuna river.

4.4. POD, FAR, and BIAS Assessment of All Simulations and Best Model Results. Both Testebo and Voxna rivers had the highest hit rate or probability of detection (POD) value of 0.98 attained for the 50 and 25 m DEMs, whereas the higher resolution DEMs received lower maximum values (Figure 9). It can also be seen in the graphs that there was more spread in the POD values of results that used finer resolution, compared with coarser resolution DEMs (due to the effect of roughness), especially for the Testebo river. Higher overall POD scores were computed for Voxna river. The models that used 10 to 50 m data for this study site received higher POD values ($POD \geq 0.886$). Eskilstuna showed an opposite trend in the probability of detection scores. Unlike the two other rivers, the maximum POD value (0.98) was attained in both sites by the 1 m resolution data. It is also shown in the figure that the POD decreased as the resolution became coarser.

If the POD scores of the most optimal results according to the different $F$-statistics (as marked in Figure 9) will be considered, it can be seen that they varied with the different goodness-of-fit measures. The most optimal results from $F_{2_O}$, had lower hit rate compared with the results from $F_{2_U}$ and $F_{1} | F_{3}$. In the case of Testebo, best models quantified using $F_{2_U}$ received the highest POD scores. For Voxna, the hit rates of most optimal results based on $F_{1} | F_{3}$ and $F_{2_U}$ were very close to each other.

False alarm ratio values have inverse performance relationship from PODs; the closer the FAR value is to 0 (maximum), the better. The pattern manifested in false alarm ratio most likely agreed in all three rivers; i.e., they decreased as the resolution became coarser. Simulations that came from the 50 m DEM also received the lowest FAR values, specifically for the Testebo river ($FAR = 0.45$). The highest FAR scores attained were as follows: Testebo = 0.0198 (10 m), Voxna = 0.037 (5 m), EskilstunaNorth = 0.014 (1 and 2 m), and EskilstunaSouth = 0.009 (1 m).

Models with the highest performance using $F_{2_O}$ had FAR values closer to the maximum. FAR values of $F_{2_U}$ results for Testebo river were farther from 0, especially for the 1 m data, while the FAR score for the 50 m DEM was closer to 0. In Voxna, FAR values of highest performing models quantified through $F_{2_U}$ and $F_{1} | F_{3}$ were the same. Like Testebo river, the highest resolution DEMs (3 m and 4 m) had lower FAR scores, while the rest had values closer to the maximum FAR score of 0.

Looking at the BIAS scores of the different simulations results, it can be seen how the lower resolution DEMs, mainly the 50 m data, produced results that overestimated the flooding. For Testebo and Voxna rivers, there were more variations in the BIAS values of the higher to medium
resolution DEMs. In the former, the results from 1 m to 10 m DEMs produced more results that underestimated the flooding \( (BIA S < 1) \), while the 25 and 50 m gave more overpredictions \( (BIA S > 1) \). For the latter, the 10 to 50 m DEM caused overestimations in the model.

The majority of the best model results derived using \( F_1 | F_3 \) and \( F_2_U \) for Testebo and Voxna rivers had \( BIA S > 1 \), especially for \( F_2_U \). For the Voxna river, the \( BIA S \) scores of the best models from \( F_1 | F_3 \) and \( F_2_U \) for the lower resolution DEMs were closer to 1. With \( F_2_O \), optimal models from almost all the resolutions for the Testebo river had \( BIA S \) towards underprediction. With Voxna river, the high-resolution DEMs with maximum \( F_2_O \) values also underpredicted the inundation. From 10 to 50 m, despite the overprediction, the \( BIA S \) scores were closer to 1. Computed \( BIA S \) scores for the Eskilstuna river were all closer to 1, ranging from 0.946 to 1.106. All DEM resolutions, with the exception of the 50 m data, had values less than 1. The 1 m DEMs had the \( BIA S \) score closest to 1 (i.e., 0.99), followed by the 25 m (0.989) DEM, in both sites.

**Figure 6**: Model performances of simulation results using different DEM resolutions for the northern and southern parts of the Eskilstuna river.

**Figure 7**: Flood extents for the Eskilstuna river using different resolution DEMs.
5. Discussion

If models are calibrated using a combination of DEM and roughness values, the best performing models varied among the feature agreement statistics used, as what was exemplified in the Testebo and Voxna cases. This was because the different equations try to handle the sizes of overlap, overestimation, and underestimation differently in the numerator. Also, as roughness values have impact on the size of the flooding produced (i.e., flood area increases with increased Manning’s $n$), the friction coefficient paired with a given DEM can affect the scores in a given performance measure, especially with $F_{O2}$ and $F_{U2}$. The most optimal models derived for different resolution DEMs using $F_{O2}$ were paired with lower roughness values (mostly from 0.01 to 0.05) (Figure 4), as they produced smaller inundations. Since the equation tries to penalise the overlap with the overprediction by the model, a lower size of overestimation will give higher performance scores. Contradictory to this was the result using $F_{U2}$, which tries to suppress the

![Figure 8: Changes in overlap and over- and underestimation sizes in the different study cases.](image-url)
underestimation produced in the model. Thus, models with lowest underestimations, i.e., those produced by higher roughness values, received maximum performance using this measure. F3, on the contrary, produced the lowest overall performance in all feature agreement statistics although the combinations of best performing models it considered were the same as F1. The equation tries to limit the effect of the overlap size to the quantified performance value, which is the problem with F1 that was mentioned by Hunter [9]. In the constraint that F3 applies in the numerator, it subtracts the overlap with both the sizes of over- and underestimation, balancing the effects of $F_2O$ and $F_2U$. This prevents the model from attaining higher scores, which can help identify better performing models. This also means that if the F3 value is high, the model can be considered to be very good. However, the use of $F_2O$ and $F_2U$ may be good to use if the modeller seeks to find minimum or maximum, respectively, probable flood inundation areas.

Furthermore, the roughness value used together with the DEM resolution was important to get higher performance in the case of Testebo and Voxna rivers. As generally seen in the results of the performance scores (Figure 3), higher roughness values worked better with higher resolution DEMs, while lower resolution DEMs performed better with lower Manning’s n. Also clear is that a particular DEM paired with the right Manning’s n can attain high performance values, which was especially true for the coarser resolution DEMs. As shown in Tables 3 and 4, the maximum performance scores received by the lowest resolution DEMs (25 and 50 m) in at least three goodness-of-fit measures (F1, $F_2O$, and F3) were higher than the finest resolution DEMs. The overlap size with the reference data produced by them was also the highest (Figure 8). Even the POD scores (Figure 9) were high and yielded the least varied values compared with the other resolution data. But, despite the high quantified performance that they received, the extents generated by them were not better than the higher resolution DEMs in terms of details and their accuracy at the edges (Figures 5 and 7). These resolutions (specifically 50 m) also produced more overestimation of the flooding, lower FAR scores, and gave higher BIAS in overestimation.

Peculiarly, the high performance scores manifested by the low-resolution DEMs for the Testebo and Voxna rivers were opposite to that of the Eskilstuna river, wherein the highest resolution DEMs received the highest performance scores. In addition, a more significant decrease in performance is noticeable for the Eskilstuna test sites as the resolution was decreased. Even when the performance scores for the Testebo and Voxna rivers using different DEMs paired with similar Manning’s n of 0.03, the 1–5 m DEMs received lower performances than the coarsest resolution DEMs. One of the possible explanations can be the reference data used to compare the results to derive the different performance values. For both Testebo and Voxna rivers, historical flood observations were used for extent validation, while for Eskilstuna, the reference data were derived from simulations using the original laser scanned point cloud data of 0.78 m (North) and 0.86 m (South) DEMs, due to the absence of a flood extent from historical flood event. The 1 m DEM gave the highest performance in all F-statistics (from $F = 0.93–0.974$ in Table 5).
and better details and match in the borders generated (Figure 7) because it was most similar to these two resolutions. The 50 m, on the contrary, was the most dissimilar in extent and performance because of the resolution difference from the reference model. The tendency of getting higher performance for the finest DEM resolution when benchmarking with a better topographic data was similar to the results of Cook and Merwade [5] and Peña and Nardi [6]. As also seen in Table 5, the percent difference in the inundated area between the reference model and the modelled floods became larger (due to the modelled areas getting smaller) as DEM resolution changed from 1 to 3 m. Why the modelled areas became smaller for 3 m than 1 m resolution is unclear to us, but it may have to do with local conditions and where the elevation data points have been sampled. However, with further change in resolution, the percent difference first decreased between 10 and 25 m DEM, followed by an increase for 50 m resolution. The modelled inundated areas are now getting bigger and bigger. The effect is most profound for the Northern area, which has a narrower floodplain. The reason for this behaviour can be attributed to raised bed levels when the DEM cell size is getting closer to or even exceeding the width of the channel. Small depressions that used to be present in the high-resolution DEMs were removed and replaced by higher lying areas, with elevated water levels as a result. With very big cell sizes, the cells would then represent the surrounding terrain, rather than being representative of the river channel. This effect was shown in Lim and Brandt [12] for the Testebo and Voxna rivers. Cook and Merwade [5] and Saksena and Merwade [4] also show in their results that the size of inundation areas become larger when using coarser grid resolution. Hence, from a flood risk perspective, the use of coarse-resolution models may actually increase the chances of being on the “safe” side when the flood map is used for planning purposes.

Maximum scores in all performances (from the entire ensemble of models) for both Testebo and Voxna rivers were also lower than for the Eskilstuna river. This is not surprising as the deviations in result for the Eskilstuna river only depended on the quality of the DEMs. However, besides adding roughness as an influencing parameter, these low scores could also have been influenced by the historical data used for the validation. According to Lim and Brandt [12], the data can have accuracy issues in terms of how they were generated (e.g., how precise the flood borders were digitised or produced), and their timeliness (i.e., the time difference between the event and generation of topographic data), which can affect getting better inundation matches. Thus, an accurate estimation of performance scores can also be subjective to the observation data, as what is stated in Merwade et al. [24]. Moreover, a perfect or very high match between the modelled and real flood extents can be difficult to achieve because of other factors, which can affect the actual flood extents, that are not taken into account by the model used or in the modelling performed.

The single roughness Manning’s n of 0.033 also worked well for Eskilstuna and Voxna rivers in the four performance estimators although this did not produce the most optimal results for the latter. This was regardless of the hydraulic models used (HEC-RAS 1D for Eskilstuna and 2D CAESAR-LisFlood for Voxna). However, for the Voxna river, performance scores among the different resolution DEMs using this friction had lower variability in the three feature agreement statistics (except for $F_{2U}$) than the Eskilstuna river. But even when considering the performance results from the ensemble, the quantified performances for the Voxna river were higher and less varied than the two other study areas. For the Testebo river, this roughness led to lower performance for higher resolution DEMs, especially when using $F_{2U}$ and $F_3$. This was for the reason that higher roughness values worked well with these resolutions for Testebo river, while lower resolution DEMs were better with lower roughness values.

The result that the Eskilstuna river’s flat terrain area received better performance measures than the steeper area might at first seem surprising, as the modelled flood borders clearly are at further distance from the reference model in the flat area. Nonetheless, Cook and Merwade [5] also achieved similar results in their study over the flat Brazos river (better performance) and the steeper Strouds creek (lower performance). This can be attributed to the flatter sites’ relatively larger inundated areas in comparison to the fringe that will be either over- or underestimated. But, it is still important to notice that the disparity distances between the modelled and the reference flood edges are always bigger in the flatter areas. This makes it questionable to only rely on the feature statistics treated in this paper. Instead, the

| Table 5: Performance scores of results from the Eskilstuna river, based on using different resolution DEMs. |
|---------------------------------------------------------------|
| Resolution (m) | $F_1$ | $F_{2U}$ | $F_{2I}$ | $F_3$ | Inundation area (km²) | % diff. |
|----------------|-------|---------|---------|-------|----------------------|---------|
| Eskilstuna north |       |         |         |       |                      |         |
| 1              | 0.965 | 0.951   | 0.944   | 0.930 | 0.1604               | −0.67   |
| 2              | 0.956 | 0.943   | 0.926   | 0.912 | 0.1588               | −1.69   |
| 3              | 0.911 | 0.892   | 0.839   | 0.821 | 0.1528               | −5.41   |
| 10             | 0.874 | 0.835   | 0.786   | 0.747 | 0.1532               | −5.16   |
| 25             | 0.814 | 0.726   | 0.716   | 0.629 | 0.1598               | −1.09   |
| 50             | 0.724 | 0.543   | 0.630   | 0.446 | 0.1786               | 10.55   |
| Eskilstuna south |      |         |         |       |                      |         |
| 1              | 0.974 | 0.966   | 0.958   | 0.949 | 0.3578               | −0.83   |
| 2              | 0.959 | 0.946   | 0.930   | 0.917 | 0.3548               | −1.68   |
| 3              | 0.943 | 0.926   | 0.904   | 0.887 | 0.3526               | −2.30   |
| 10             | 0.920 | 0.886   | 0.875   | 0.841 | 0.3567               | −1.15   |
| 25             | 0.847 | 0.776   | 0.765   | 0.694 | 0.3567               | −1.15   |
| 50             | 0.754 | 0.613   | 0.648   | 0.507 | 0.3754               | 4.05    |
distance disparities described in Brandt [21] and Lim and Brandt [12] should also be considered. These measures will provide better estimates on local flood extent uncertainties, especially where the terrain is flat.

Finally, it is clear from the results that when high-resolution DEMs are used, calibration of the roughness coefficient is of utmost significance. Important to note here that calibrated results from earlier models, as well as standard Manning’s n values, are not advisable to be used. Calibration should be done according to the current DEM, study area, model, and model conditions. As what is shown in Figure 4, although the general trend for roughness values and DEM for each performance measure have similarities, the most optimal roughness still differed. These differences in the friction values reflect sampling variation, which is typical for any model parameter. In the case of Testebo and Voxna, the varying roughness values can be ascribed to the differences in their physical characteristics, which affect the flow characteristics in the river. As different study areas are spatially heterogeneous, the friction values will vary from one site to another. In addition, because the DEM provides the topography of the area for the model, its quality in terms of resolution will affect the flow characteristics, which in turn can impact the friction to be used and the flood extent produced. The usage of the 1D or 2D model can also have an effect on the optimal roughness values derived, due to their assumptions in solving the energy loss equation [25] (cf. Horritt and Bates [7]). For the 2D model results (Voxna and Testebo), if the basis of optimality will be F1 | F3 for the highest resolution DEMs (1–5 m), the most optimal Manning’s n is 0.08 but varied as the resolution decreased. For the Eskilstuna river, the most optimal roughness will be difficult to know as only a single roughness was used. Likewise, the discharge used can also produce varying inundation extents and patterns, as well as lead to different optimal roughnesses [7]. Maximum discharges were used for the different study sites. Although they were roughly of the same magnitude, the flow differed a bit (360 m³/s for Voxna, 160 m³/s for Testebo, and 198 m³/s for Eskilstuna). This can be another reason for the varying friction values. All these factors can affect the modelling results, as what is also shown in earlier studies, and as a result of the equifinality problem [1]. However, since our study focused on the variability of performance measures as effect of the DEM resolution and roughness, our results are centred to the effects of topography in the flood extents and performance measures, rather than the model or the flow used.

From a feature statistics point-of-view, it may seem unnecessary to use high-resolution DEMs, when average resolution DEMs provide similar F-statistics scores, as in the case of Voxna and Testebo rivers. However, as the results from Eskilstuna river show (wherein only the effect of the DEM is considered), feature statistics clearly get better when resolution is increased. The positive effect of resolution in the flood model results when considering both the effects of DEM and roughness became more evident when the disparity distance measures were used, especially for Voxna river. In Lim and Brandt [12], the mean and median disparities for 3 to 5 m were 59.71–63.16 m and 10.75–15.18 m, respectively. These were clearly better than for the 50 m DEM (D=117.7 m and D=41.17 m). Nonetheless, the results of the disparity measures are more unclear for the Testebo river.

6. Conclusions

This study has highlighted the implications of using feature statistics in assessing modelled flood inundation areas. Some of the most frequently used feature statistics have been tested on three different rivers, whereof one constituted two different topographies and flat and steep side slopes, respectively. From the results, it is clear how the performance values using the feature statistics varied when going from the most commonly used F1 that divides the overlap area of modelled and reference flood with the total combined area of modelled and reference flood, over the F2o and F2U that penalises over- and underpredicted areas, respectively, to F3 that penalises both over- and underpredicted areas. This suggests that a high F3 value, i.e., close to 1, more or less guarantees a high match between modelled and reference flood area. However, even when feature statistics show high values, caution should be taken especially for local or flat areas. Consequently, alternative performance measures, such as disparity distances, may be better substitutes.

The direct relationship between the roughness value and the size of flooding was also shown to affect quantified performance, especially from F2o and F2U equations. As the flooded area increases with higher Manning’s n, the tendency to overestimate flood becomes higher. Because F2o penalises the model with overestimation produced, lower roughness values that generate smaller extents and minimal overprediction performed better with this measure. On the contrary, F2U worked well with higher friction because the equation tries to minimise the underestimation made by the model.

The results also showed that DEMs of poor resolution can receive relatively high performance scores over a wide range of roughnesses. But despite the high scores, there are more spatial inconsistencies in the flood extents produced from low-resolution data. Therefore, assumptions on the goodness of results based on poorer resolution DEMs must be carefully made. Usage of higher resolution DEMs remains advantageous for modelling as they represent the topography of the study area better. As the results from the Eskilstuna river showed, which is only dependent on the quality of the DEM, better resolution has the potential to always bring better results. However, high-resolution DEMs needs to be carefully calibrated through the roughness value, to be able to fully utilise their modelling advantage. If not calibrated properly, the feature agreement statistics will directly show decreased scores.

Performance scores also varied in the different study areas. Regardless of the measure used, Eskilstuna and Voxna rivers received higher overall performances than the Testebo river. These variations in the scores can be caused by several factors such as the modelling assumptions of the model utilised (1D vs. 2D) or the spatial heterogeneities of the study areas, which can both influence the flow being modelled, and
the inundation extents produced. Moreover, the reference data and its quality can also affect the quantified performance.

**Data Availability**

The data used to support the findings of this study have not been made available because they are the property of Lantmäteriet (the Swedish mapping, cadastral, and land registration authority) and the municipalities of the studied areas.

**Disclosure**

This research has been carried out as part of the employment at University of Gävle.

**Conflicts of Interest**

The authors declare that there are no conflicts of interest regarding the publication of this paper.

**Acknowledgments**

We would like to thank the reviewers for constructive feedback on a previous version of this paper.

**References**

[1] K. Beven and J. Freer, "Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology," *Journal of Hydrology*, vol. 249, no. 1–4, pp. 11–29, 2001.

[2] N. J. Lim, *Modelling, Mapping and Visualisation of Flood Inundation Uncertainties*, Ph.D. thesis, University of Gävle, Gävle, Sweden, 2018.

[3] P. D. Bates and A. P. J. de Roo, "A simple raster-based model for flood inundation simulation," *Journal of Hydrology*, vol. 236, no. 1–2, pp. 54–77, 2000.

[4] S. Saksena and V. Merwade, "Incorporating the effect of DEM resolution and accuracy for improved flood inundation mapping," *Journal of Hydrology*, vol. 530, pp. 180–194, 2015.

[5] A. Cook and V. Merwade, "Effect of topographic data, geometric configuration and modeling approach on flood inundation mapping," *Journal of Hydrology*, vol. 377, no. 1–2, pp. 131–142, 2009.

[6] F. Peña and F. Nardi, "Floodplain terrain analysis for coarse resolution 2D flood modeling," *Hydrology*, vol. 5, no. 4, p. 52, 2018.

[7] M. S. Horritt and P. D. Bates, "Evaluation of 1D and 2D numerical models for predicting river flood inundation," *Journal of Hydrology*, vol. 268, no. 1–4, pp. 87–99, 2002.

[8] M. S. Horritt and P. D. Bates, "Predicting floodplain inundation: raster-based modelling versus the finite-element approach," *Hydrological Processes*, vol. 15, no. 5, pp. 825–842, 2001.

[9] N. Hunter, *Development and Assessment of Dynamic Storage Cell Codes for Flood Inundation Modelling*, Ph.D. thesis, University of Bristol, Bristol, UK, 2005.

[10] G. Aronica, P. D. Bates, and M. S. Horritt, "Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE," *Hydrological Processes*, vol. 16, no. 10, pp. 2001–2016, 2002.

[11] V. Tayefi, S. N. Lane, R. J. Hardy, and D. Yu, "A comparison of one- and two-dimensional approaches to modelling flood inundation over complex upland floodplains," *Hydrological Processes*, vol. 21, no. 23, pp. 3190–3202, 2007.

[12] N. J. Lim and S. A. Brandt, "Flood map boundary sensitivity due to combined effects of DEM resolution and roughness in relation to model performance," *Geomatics, Natural Hazards and Risk*, vol. 10, no. 1, pp. 1613–1647, 2019.

[13] K. J. Beven, *Environmental Modelling: An Uncertain Future*, CRC Press, Boca Raton, FL, USA, 2009.

[14] A. H. Murphy, "What is a good forecast? An essay on the nature of goodness in weather forecasting," *Weather and Forecasting*, vol. 8, no. 2, pp. 281–293, 1993.

[15] I. B. Mason, "Binary events," in *Forecast Verification: A Practitioner’s Guide In Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson, Eds., pp. 37–76, John Wiley & Sons, West Sussex, UK, 2003.

[16] J. Schaefer, "The critical success index as an indicator of warning skill," *Weather Forecast*, vol. 5, no. 4, pp. 570–575, 1990.

[17] N. M. Hunter, P. D. Bates, M. S. Horritt, A. P. J. De Roo, and M. G. F. Werner, "Utility of different data types for calibrating flood inundation models within a GLUE framework," *Hydrology and Earth System Sciences*, vol. 9, no. 4, pp. 412–430, 2005.

[18] D. C. Mason, P. D. Bates, and J. T. Dall’ Amico, "Calibration of uncertain flood inundation models using remotely sensed water levels," *Journal of Hydrology*, vol. 368, no. 1–4, pp. 224–236, 2009.

[19] D. Klang and K. Klang, "Analyse Av Höjdmodeller för Översvämningssimulering," Unpublished Report for Myn-digheten för samhällsskydd och beredskap and Lantmäteriet, October 2009, https://docplayer.se/docview/57/39936952.pdf.

[20] S. A. Brandt, "Betydelse av höjdmodellers kvalitet vid endi-mensionell översvämningssimulering," Höskolan i Gävle, Gävle, Sweden, FoU-Rapport Nr. 35, 2009.

[21] S. A. Brandt, "Modeling and visualizing uncertainties of flood boundary delineation: algorithm for slope and DEM resolution dependencies of 1D hydraulic models," *Stochastic Environmental Research and Risk Assessment*, vol. 30, no. 6, pp. 1677–1690, 2016.

[22] T. J. Coulthard, J. C. Neal, P. D. Bates, J. Ramirez, G. A. M. de Almeida, and G. R. Hanrock, "Integrating the LISFLOOD-FP 2D hydrodynamic model with the CAESAR model: implications for modelling landscape evolution," *Earth Surface Processes and Landforms*, vol. 38, no. 15, pp. 1897–1906, 2013.

[23] US Army Corps of Engineers Hydrologic Engineering Center, *HEC-RAS: River Analysis System. User’s Manual (v. 4.0)*, US Army Corps of Engineers Hydrologic Engineering Center, Davis, CA, USA, 2008.

[24] V. Merwade, F. Olivera, M. Arabi, and S. Edleman, "Uncertainty in flood inundation mapping: current issues and future directions," *Journal of Hydrologic Engineering*, vol. 13, no. 7, pp. 608–620, 2008.

[25] N. M. Hunter, P. D. Bates, M. S. Horritt, and M. D. Wilson, "Simple spatially-distributed models for predicting flood inundation: a review," *Geomorphology*, vol. 90, no. 3–4, pp. 208–225, 2007.
