Vanmaercke Matthias (Orcid ID: 0000-0002-2138-9073)
Chen Yixian (Orcid ID: 0000-0002-7710-580X)
Campforts Benjamin (Orcid ID: 0000-0001-5699-6714)

Status: 04 September 2020
Journal: Earth Surface Processes and Landforms

Predicting gully densities at sub-continental scales: a case study for the Horn of Africa

Running head title: Predicting gully densities at sub-continental scales

Matthias Vanmaercke[1,*], Yixian Chen[1,2], Nigussie Haregeweyn[3], Sofie De Geeter[1,4], Benjamin Campforts[5], Wouter Heyndrickx[6], Atsushi Tsunekawa[7], Jean Poesen[4,8]

[1] Université de Liège, Département de Géographie, Liege, Belgium
[2] Institute of Soil and Water Conservation, Chinese Academy of Science and Ministry of Water Resources, Yangling, Shaanxi, China
[3] International Platform for Dryland Research and Education, Tottori University, Tottori 680-0001, Japan
[4] University of Leuven, Department of Earth and Environmental Sciences, Division of Geography and Tourism, Leuven, Belgium
[5] CSDMS, Institute for Arctic and Alpine Research, University of Colorado at Boulder, Boulder, CO, USA
[6] Independent Scholar, Edegem, Belgium
[7] Arid Land Research Center, Tottori University, 1390 Hamasaka, Tottori 680-0001, Japan
[8] Faculty of Earth Sciences and Spatial Management, UMCS, Lublin, Poland

[*] Corresponding Author: matthias.vanmaercke@uliege.be

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/esp.4999

Abstract

Despite its environmental and scientific significance, predicting gully erosion remains problematic. This is especially so in strongly contrasting and degraded regions like the Horn of Africa. Machine learning algorithms like Random Forests (RF) offer great potential to deal with the complex, often non-linear nature of factors controlling gully erosion. Nonetheless, their applicability at regional to continental scales remains largely untested. Moreover, such algorithms require large amounts of observations for model training and testing. Collecting such data remains an important bottleneck.

Here we help addressing these gaps by developing and testing a methodology to simulate gully densities across Ethiopia, Eritrea and Djibouti (total area: 1.2 million km²). We propose a methodology to quickly assess the gully head density (GHD) for representative 1-km² study sites by visually scoring the presence of gullies in Google Earth and then converting these scores to realistic estimates of GHD. Based on this approach, we compiled GHD observations for 1,700 sites. We used these data to train sets of RF regression models that simulate GHD at a 1 km² resolution, based on topographic/geomorphic, land cover, soil and rainfall conditions. Our approach also accounts for uncertainties on GHD observations. Independent validations showed generally acceptable simulations of regional GHD patterns. We further show that: (i) model performance strongly depends on the amount of training data used; (ii) large prediction errors mainly occur in areas where also the predicted uncertainty is large; and (iii) collecting additional training data for these areas results in more drastic model performance improvements. Analyses of the feature importance of predictor variables further showed that patterns of GHD across the Horn of Africa strongly depend on NDVI and annual rainfall, but also the normalized steepness index (k_{sn}) and distance to rivers. Overall, our work opens promising perspectives to asses gully densities at continental scales.

Keywords: Gully erosion; Ethiopia; Eritrea; Djibouti; Google Earth; Land Degradation; Random Forests; Arid region

1. Introduction

Gully erosion is a major concern in many regions worldwide (e.g. Poesen et al., 2003; Valentin et al., 2005). Especially in arid and semi-arid areas, gullies can be a major cause of land degradation (Vanmaercke et al., 2011). At the hillslope scale, they often lead to important soil losses, direct losses of productive land and reduced biomass production (e.g. Avni 2005; Frankl et al., 2016). At the catchment scale, gullies are a dominant sediment source and can significantly increase catchment connectivity (de Vente and Poesen, 2005; Vanmaercke et al., 2012; de Vente et al., 2008), resulting in negative downstream impacts like reduced water quality and reservoir storage capacity losses (Owens et al., 2005; Haregeweyn et al., 2005; Vanmaercke et al., 2011). Likewise, gullies can lead to higher flood frequencies and magnitudes (e.g. Martineli Costa et al., 2007).

One region that is heavily affected by these problems is the Horn of Africa and, in particular, the northern part of the East-African Rift (i.e. Ethiopia, Eritrea and Djibouti). While this region is considered to be one of the "water towers" of Africa (e.g. Hamond, 2013; Zenebe et al., 2013), it is also one of the main hotspots in terms of soil erosion (Borrelli et al., 2017; Fenta et al., 2020) and catchment sediment export (Vanmaercke et al., 2014). This strongly limits the possibility to use available soil and water resources in a sustainable way (e.g. Haregeweyn et al., 2005; 2006; Rosa et al., 2020). Increasing land use pressure (e.g. Borrelli et al., 2017) and climate change (e.g. Pelletier, 2015; Li and Fang, 2016; Vanmaercke et al., 2016) are likely to further aggravate these challenges. Coordinated efforts are therefore required. Over the past decades, soil and water conservation measures have been implemented at a large scale in the Horn of Africa (Nyssen et al., 2004; Hargeweyn et al., 2015) and significant progress has recently been made in assessing the overall susceptibility of this region to sheet and rill erosion (Haregeweyn et al., 2017; Fenta et al., 2020) and landsliding (e.g. Broeckx et al., 2018; 2020). However, no such tools exist for assessing gully erosion. Current efforts to identify gully erosion hotspots mainly rely on simplified, expert-based assessments that are based on few observations and generally remain unvalidated (e.g. Haregeweyn et al., 2017).

The inability to assess gully erosion risks over larger areas not only affects the Horn of Africa, but relates to a more general and fundamental challenge. Despite considerable research attention over the past decades (e.g. Torri and Poesen 2014; Castillo and Gomez, 2016; Vanmaercke et al. 2016) our ability to simulate gully erosion remains very limited, especially at regional to continental scales (e.g. de Vente et al., 2013; Poesen, 2018). Several process-oriented models have been proposed (e.g. Poesen et al. 2011; Campo-Bescos et al.,

2013; Torri et al. 2018; Sidorchuk 2020) but these models typically have very high input data requirements and are practically impossible to apply over larger areas (e.g. Merrit et al., 2003; Poesen et al., 2011; de Vente et al., 2013). Assessing gully erosion at regional to sub-continental scales therefore needs to resort to more simplified, empirical approaches. However, very few studies have currently attempted to do this (e.g. Poesen, 2018).

Empirical approaches also come with their own challenges and limitations. Overall, gully erosion can be controlled by a wide range of environmental conditions, which may vary both locally and at regional scales. These include topographic conditions (e.g. slope and contributing area; Torri & Poesen, 2014; Vanmaercke et al., 2016), vegetation characteristics (e.g. Zhao et al., 2016; Vannoppen et al., 2015), lithology and soil characteristics (e.g. Radoane et al., 1995; Knapen et al., 2007), weather conditions (e.g. Thompson, 1964; Ionita, 2006; Vanmaercke et al., 2016; Hayas et al. 2017a) and potentially tectonic activity (e.g. Menéndez-Duarte et al., 2007; Cox et al., 2010). Assessing gully erosion across large and contrasting regions therefore requires tools that account for all pertinent variables. However, as other studies indicated (e.g. de Vente et al., 2011; Vanmaercke et al., 2014; Golosov et al., 2018), classical multiple regression analyses quickly become unsuitable for such purpose. To a large extent, this is because such methods generally cannot fully account for the many (often non-linear) interactions that may exist between variables. More specifically, multiple regression typically assumes that relations between dependent and independent variables are valid over the whole domain of observations. In reality this is not necessarily so. For example, in the case of gully erosion, the topographic conditions leading to gully initiation will also depend on local soil and vegetation characteristics and the overall geomorphic setting (e.g. Torri and Poesen, 2014; Rossi et al., 2015; Amare et al., 2019).

Methodological tools are needed that allow identifying and accounting for such complexities. Advances in machine learning open promising perspectives in this regard (e.g. Youssef et al., 2016; Rahmati et al., 2017; Gayen et al., 2019). Random Forests (RF; Breiman, 2001) are a commonly applied and often successful approach (e.g. Chen et al., 2017; Hosseinalizadeh et al., 2019). Overall, a RF algorithm is based on the construction of an ensemble of decision trees that subdivide observations in groups that show maximum similarity within groups and dissimilarity between groups at every node in a tree. This is done by creating binary splitting rules, based on one predictor variable at every node (e.g. Louppe, 2014). Different decision trees are trained based on random subsets of observations. At each node, also random subsets of all the potential predictor variables are considered to determine the optimal split. Given the variation in observations and predictor variables between the trees, each tree will likely differ

and capture different aspects of the dataset. Strong patterns are likely to be captured by many trees, while weak patterns will only occur in few trees. By grouping these different trees and looking at their majority outcome (in the case RF classification) or their average assigned value (in the case of RF regression), one can obtain robust and accurate predictions (e.g. Louppe, 2014). A major advantage of RF algorithms is that they can deal well with non-linearity and inter-correlations between variables. They do not presuppose uniform relationships across the variable space, nor that these variables follow a specific distribution. Furthermore, the complexity of random forests can be adapted in function of the amount of training data available, through the tuning of hyperparameters such as tree depth and the number of trees (Breiman, 2001; Louppe, 2014). Given that gully initiation is often a non-linear, threshold-dependent process, depending on a wide range of interacting factors, RFs offer great potential to characterize this complexity.

Indeed, several studies have already applied RFs to predict the occurrence of gully heads with significant success (e.g. Gayen et al., 2019; Rahmati et al., 2017; Arabameri et al., 2018). However, nearly all of these studies focused on smaller study areas (i.e. < 1,000 km²). Its applicability at regional to sub-continental scales (e.g. 100,000 – 1,000,000 km²) remains largely untested. Nonetheless, Jurchescu and Grecu (2015) showed that uncertainties of classification and regression trees are clearly higher when applied over larger areas. In addition, most studies so far focused on using RF classification to predict the absence or presence of gullies at the pixel scale (e.g. Gayen et al., 2019; Rahmati et al., 2017; Arabameri et al., 2018). The potential of RF to simulate gully density (GD) in a continuous way remains largely unknown.

Given the nature of the algorithm, the reliability of RF applications can strongly depend on the data used to train the model. Robust yet accurate predictions typically require large amounts of data in order to avoid overfitting (e.g. Louppe et al., 2014). Likewise, applying a RF to data other than it was trained for can result in large uncertainties. As such, RFs face a similar constraint as process-based models: simulating the occurrence of gullies at regional to continental scales requires large amounts of sufficiently detailed input data. The increasing availability of numerous remote sensing products now make it possible to characterize relevant environmental factors at such scales (Vanmaercke et al., subm.). Nevertheless, accurate inventories of mapped gullies, required for model training and validation, remain difficult and labour-intensive to construct, especially in the Global South (e.g. Mararakanye et al., 2012; Golosov et al., 2018; Guyassa et al., 2018; Kariminejad et al., 2019; Yibeltal et

al., 2019). However, freely available tools like Google Earth open up promising perspectives in this regard (e.g. Zhao et al., 2016; Karydas and Panagos, 2020).

The goal of this paper is therefore twofold. Our direct goal is to develop the first data-driven gully density map at a 1 km² resolution for the Horn of Africa (i.e. Ethiopia, Eritrea and Djibouti; covering an area of around 1.2 million km²). On a more general and fundamental level, we aim to develop and test a methodology that allows assessing gully densities at regional to continental scales with a reasonable amount of effort and using freely available data sources. To this end we propose a novel gully density mapping procedure, using a random sampling strategy of observation sites for which gully densities are estimated in a fast and coherent way. These gully density data are then used to train a set of RF regression models that simulate spatial patterns of gully density while accounting for observation errors on the training data.

2. Materials and methods

2.1 Gully head density as a proxy for gully occurrence

Assessing gully densities at regional to continental scales first requires deciding on a measure to quantify these densities. Previous studies have mainly mapped gullies as linear features, expressing their occurrence as a linear density (e.g. Golosov et al., 2018) or as polygons, expressing their density as an areal fraction (e.g. Mararakanye et al., 2012). Such mapping efforts are typically very time consuming. Alternatively, Zhao et al. (2016) approximated areal gully densities at the catchment scale by assessing which fraction of a set of randomly distributed points were located within a gully. While considerably faster, such approach also comes with larger uncertainties and an important loss of information.

In this study we opted for an alternative proxy, i.e. the gully head density (GHD) which is the number of gully heads per square kilometre [# km^{-2}]. This measure offers several advantages. Practically, mapping gully heads is generally faster and more straightforward than mapping gullies as lines or polygons. For example, the distinction between a gully channel and an (ephemeral) river is often hard to make, especially in semi-arid environments. Also, from a geomorphic perspective, the distinction is largely arbitrary (e.g. Nachtergaele et al., 2002; Poesen et al. 2003). Nonetheless, whether or not a channel is mapped as a gully can have a large impact on the resulting linear or areal GD. Such interpretation difficulties are largely avoided when using GHD as a proxy. Furthermore, although gullies can easily obtain lengths

of several hundreds to thousands of meters (e.g. Dube et al., 2020), it are often the local conditions at the gully heads that control their formation and evolution (e.g. Torri & Poesen, 2014). Focussing on gully heads allows for a closer proximity between the studied phenomenon and its controlling factors, making empirical analyses more straightforward. Finally, while gullies can also widen and deepen, most expansion of gullies typically happens through headcut retreat (e.g. Vanmaercke et al., 2016; Hayas et al., 2017b). As such, it can be expected that GHD provides a stronger indicator for actual gully erosion risks than areal or linear GD.

Some evidence for this is also provided in Figure 1. For a set of 15 catchments in Ethiopia (ranging in size between 0.7 and 975 km²) for which average sediment yield measurements (SY) were available (Vanmaercke et al., 2014), we quantified the average gully densities in two different ways. First, we estimated the areal fraction of each catchment that is gullied, using a sampling procedure in Google Earth as proposed by Zhao et al. (2016). Second, we estimated the number of gully heads in each catchment using the same imagery and a similar procedure as the one explained below (section 2.2). While both proxies are positively correlated to SY, GHD clearly shows a stronger correlation. Both proxies are also inter-correlated, but only to a limited degree (Figure 1c). Given the limited number of catchments and the rather rudimentary nature of this comparison, these findings should evidently be interpreted with caution. Nonetheless, it does provide further indication that GHD is a more relevant proxy for gully erosion risk (and hence SY) than the areal fraction of gullies.

2.2 An effective strategy to map gully densities at sub-continental scales

Given their extent and the large number of gullies present, constructing complete databases of mapped gully heads at regional to continental scales remains practically unfeasible. Nonetheless, simulating gully densities at such scales requires a representative and sufficiently detailed training dataset (cf. section 1). We therefore developed a procedure where we quantified the GHD for a number of randomly selected observation sites across the study region.

Each observation site consists of a square of 1x1 km², providing a good trade-off between level of detail and robustness of the obtained GHD value. It also corresponds to the intended resolution of our GHD map. The position of each site was determined randomly with the only restrictions that a sufficiently detailed and clear Google Earth image was available and that

less than 50% of the area was covered by water. In cases where multiple images were available, the most suitable image was selected. This was decided based on the overall resolution and sharpness of the image, its colour contrasts and potential shadows. Furthermore, an overall preference was given to images taken during periods of low vegetation cover. When multiple suitable images were available, we generally used the most recent one. When no suitable image was available, the site was not considered.

Figure 2 provides an example of an observation site, demonstrating how GHD was assessed. Each site was subdivided into nine cells of equal size (ca. 0.11 km²). By zooming in to a mapping height of ca. 500 meters, the presence or absence of gully heads was assessed for each of these cells (starting in the NW corner and following the order indicated by the white numbers in Figure 2). All mapping was conducted by people with a sufficient training in geomorphology and the process of gully erosion. When evaluating whether a cell contained gully heads, we took into account the geomorphic context of the features. For example, gullies usually form on relatively steep slopes and/or have a significant contributing area. Likewise, they tend to follow the steepest slope. Furthermore, gully heads often occur in groups or as part of a dendritic gully network (e.g. insets in Figure 2). As a criterion, we interpreted a given point as an individual gully head if the gully channel length from that point to the outlet or confluence with another channel was at least 10 m and larger than the width of the channel.

Actually mapping or even counting individual gully heads in an observation site remains very labour intensive (e.g. De Geeter et al., 2019). We therefore made a crude assessment of the overall number of gully heads per site, which is fairly straightforward and much faster. For this, we assigned a score to each cell (Figure 2). Cells with no visible gully heads received a score of "0". Cells with 1-5 gully heads received a score of "1", those with 6-20 gully heads a score of "2" and those with >20 gully heads a score of "3". This approach drastically reduced the time required to assess the GHD of an observation site (on average, 1-2 minutes), but also induces a loss of information. Nevertheless, for our purposes, having counted and mapped each individual gully head would only provide a limited added value, given (i) the resolution and size of our intended GHD map; (ii) the stochastic nature of our modelling approach, but also the formation of gully heads (e.g. Hayas et al., 2017a); and (iii) the interpretation difficulties and errors that would also be associated with mapping individual gully heads (e.g. Maugnard et al., 2014).

Hence, for each site, we obtained nine scores (Figure 2). Each score was then converted into a possible number of gully heads, based on an earlier developed database of 1 km² study sites across Africa and Europe (n=2014) for which the individual gully heads were actually mapped (e.g. De Geeter et al., 2019). Each of these mapped sites was subdivided into nine cells (of the same dimensions as the cells in this study) and the number of gully heads within each cell was counted. From these counts, we derived cumulative probability distributions of the number gully heads, corresponding to each score (Figure 3). Using these distributions, we assigned a possible number of gully heads to each cell. This was done by randomly sampling a number from a uniform distribution between zero and one and selecting, from the distribution matching with the cell score, the maximum number of gully heads that had a cumulative probability smaller than or equal to that random number. By adding up the randomly selected numbers of gully heads for the nine cells, we obtained a realistic estimate of the number of gully heads within the observation site. Evidently, for cells with a score of '0' (e.g. cells 5 and 6 in Figure 2), the assigned number of gully heads was zero. Next, we added a random number of gully heads to the observation site's total. This number was sampled randomly from a normal distribution with an average of zero and a standard deviation of 1.5 and then rounded to the nearest integer value. For sites where this resulted in a negative value, the total gully head count was set to zero. Overall, this additional random number allowed to account for mapping uncertainties that are not represented by the cumulative gully head distributions (Figure 3). Examples include uncertainties relating to whether or not certain features are actually gully heads or the fact that ephemeral gully heads may only be visible on some images. Especially for sites with low scores (e.g. all '0' or only one '1' score), accounting for these uncertainties can be important.

For each observation site, this procedure of converting assigned scores to possible numbers of gully heads was repeated 100 times. This resulted in a set of 100 possible GHDs for each site. In the rest of the text, we will refer to them as the set of 'possible observations' of each site (reflecting the fact that they are not exact values, but nonetheless based on visual observations). We will refer to the average of these 100 possible GHDs as the 'average observed GHD' for that site.

In total, we assessed the GHD for 1700 sites across the Horn of Africa (Figure 4). The dates of the images used to assess GHD ranged between 2001 and 2020. For around 90% of the sites, GHD assessments are based on imagery from the period 2010-2020.

As the availability of sufficient and representative training and validation data is a key concern (cf. section 1), we explored how including more observation sites influenced model performance. For this purpose, the compiled database consisted of four subsets. A first 'Basic' subset of 500 sites was used for training a first version of the model. An independent 'Validation' dataset (n=400) was used to test the performance of all trained models. We expanded these subsets with two additional subsets: one containing 400 more sites that were randomly distributed across the entire study area (cf. 'Random Extension'; Figure 4) and one containing sites that were randomly positioned within key areas (cf. 'Targeted Extension'). These key areas were identified by applying a first version of the random forest model (trained with the 'basic' set, according to the procedures described in section 2.4) and then selecting the zones for which the range in predicted GHDs was larger than 30 heads km^{-2}. This targeted extension was constructed to see if training the model with additional data from areas where GHD is difficult to predict, leads to stronger increases in model performance. It provides a simple example of 'active learning', with pool-based sampling and a single iteration (Settles, 2009). Overall, such strategy may limit the assessment of GHD for sites that would be uninformative to the model.

2.3 Considered predictor variables

For each observation site, a number of variables were extracted that potentially explain differences in GHDs (Table 1). These variables describe the topography and geomorphic context, land cover, soil characteristics and climate at each site.

Concerning topography and the geomorphic context, we considered the average and maximum slope of the study area, the average elevation and the profile curvature. These four variables were originally derived from 90m resolution SRTM data and directly extracted from Amatulli et al. (2018) at a 1 km² resolution. We also calculated a normalized steepness index (k_{sn}; [$m^{0.9}$]), using HYDROSHEDS data (Lehner et al., 2013). Following an approach similar to Wobus et al. (2006), we calculated the k_{sn} of each pixel at a 90m resolution as:

$$k_{sn} = S \times A^{\theta} \quad \text{(Eq. 1)}$$

With S [m/m] the slope steepness of each pixel, A [m^2] the total area draining to the pixel and θ the concavity constant. In accordance with many other studies (e.g. DiBiasi & Whipple, 2011) we set θ to 0.45. For each study site, the median k_{sn} of all pixels was then calculated. While k_{sn} is commonly used to quantify the steepness of river profiles, it also provides a

topographic proxy of flow shear stress that can occur at any pixel. Furthermore, this equation shows clear similarities with proposed equations of slope-area thresholds of gully initiation (Torri and Poesen, 2014). As such, it can be expected that high k_{sn} values correspond to higher gully occurrence probabilities. Nevertheless, gullies can also result from regressive erosion of river systems (e.g. Mendez-Duarte et al., 2007). Furthermore, they frequently occur in alluvial deposits as a result of subsurface or saturation excess overland flow (e.g. Amare et al., 2019). We therefore also included the average distance of each site to a river as a potential predictor variable. These distances were calculated using the HYDROSHEDS dataset (Lehner et al., 2013), considering river channels with a Strahler order of 4 or more.

Regarding land use and land cover, we calculated the long-term (1999-2017) average Normalized Difference Vegetation Index (NDVI) for every site. NDVI observations were derived from the Copernicus Global land Service (Copernicus Service Information, 2019) at their original resolution of 1000m. However, large parts of the study area are characterized by large seasonal contrasts in rainfall and vegetation cover (e.g. Nyssen et al., 2005), leading to potentially important interactions with soil erosion (e.g. Vanmaercke et al., 2010; Lemma et al. 2018). To account for such interactions, we also incorporated a rainfall-weighted version of the average NDVI. Using the same data source and time period, we first calculated long-term average monthly NDVI values. These values were then multiplied by their corresponding estimated average monthly rainfall (derived from Huffman et al., 2019) and added up. This sum was then divided by the total annual rainfall. Furthermore, we included a binary variable (CL), indicating whether the dominant land cover of the site was cropland or not (based on Buchhorn et al., 2019). This variable was expected to account for potential effects of land use/land management. While various soil and water conservation measures (e.g. stone bunds, soil trenches, grassed lynchets) are widely implemented across Ethiopia (e.g. Hargeweyn et al., 2015; Taye et al. 2013; 2017), no comprehensive spatial databases of these measures exist. However, since they are mainly implemented on cropland, it was expected that this variable could provide an indication on the potential effect of conservation structures on gully head development (e.g. Monsieurs et al. 2015).

Soil characteristics were derived from the recently developed SoilGrids database (Hengl et al., 2017). We extracted the estimated average soil depth, soil bulk density, volumetric rock fragment content as well as the mass percentage of clay, silt and sand. For all characteristics (except the average soil depth), we used values estimated at a depth of 5 cm.

To characterize rainfall conditions, we extracted the estimated (1979-2016) average annual rainfall (as derived by Broeckx et al., 2020 based on Beck et al., 2019) as well as two proxies

of rainfall intensity: the 99% percentile of daily precipitation for the period 1979-2016 (P_{99d}; as derived by Broeckx et al., 2020 based on Beck et al., 2019) and the Rainy Day Normal (RDN, i.e. the average rainfall depth on a rainy day). Overall, rainfall intensity is commonly recognized as a key factor driving gully erosion (e.g. Poesen et al., 2003; Vanmaercke et al., 2016; Hayas et al., 2017).

2.4 Predicting gully densities using random forests

The extracted variables (Table 1) and possible GHD values (cf. section 2.2) were used to train random forest regression models that simulate GHD. We implemented our approach using '*RandomForestRegressor*' module of the freely available scikit-learn library (Pedregosa et al., 2011; version 0.21.2) through Python version 3.7.3. Prior to training and applying our models, we conducted an exploration to identify suitable values for two hyperparameters, i.e. the maximum tree depth and number of trees in our random forest. In accordance with the bias-variance trade-off (Geurts, 2002), tree depths that are too limited will result in poor model performances due to underfitting, while overly complex trees risk being overfitted to the training data and may induce significant biases. An analogous argument can be made for the number of trees. Hence, we explored optimal values for these two hyperparameters by first randomly selecting a set of observation sites (containing 400-1000 sites that were picked from all subsets, except for the targeted extension; cf. section 2.2 and Figure 4) and training a RF model with a preselected tree depth and number of trees (using the average observed GHD of each site, cf. section 2.2). Next, we applied the model to an independent dataset (containing 200-500 sites that were also randomly selected from the same subsets but excluding sites already used for training). We then evaluated the model performance, based on the Nash-Sutcliffe Model Efficiency (Nash and Sutcliffe, 1970; cf. section 2.5) and the total bias in predictions. The latter was calculated as:

$$Bias = \frac{\sum_{i=1}^{n} O_i - \sum_{i=1}^{n} P_i}{\sum_{i=1}^{n} O_i} \qquad \text{(Eq. 2)}$$

With n the number of observation sites in the validation set, O_i the average observed GHD (cf. section 2.2) and P_i the GHD predicted by the random forest model. Bias-values <-0.1 or >0.1 were interpreted as an indication of overfitting or underfitting. We repeated this procedure for a range of maximum tree depths (between 2 and 30) and number of trees (between 5 and 100). Overall, we found that model performance did not vary strongly, and no

significant bias was present for maximum tree depths between 7 and 30. Likewise, as soon as more than 20 trees were used, this hyperparameter had little influence on the overall performance. Also taking into account computation times, we therefore set the maximum tree depth to 15 and the number of trees to 30.

We trained all further discussed RF regression models with these values, keeping the other parameters of the '*RandomForestRegressor*' function to their default setting. All variables listed in Table 1 were consistently included as predictor variables, since they each can be expected to make a meaningful contribution to explaining regional variability in GHD. While the RF algorithm and predictive performance are largely unaffected by intercorrelations between variables, intercorrelation does affect the interpretation of feature importances for RF. This is something we will keep in mind later in this paper (cf. section 4.2). To evaluate the effect of training data on model performance, we trained four different sets of RF models using four different training datasets (cf. Figure 4): (i) the basic set (n=500), (ii) the basic set plus the random extension (n=900), (iii) the basic set plus the targeted extension (n=900), and (iv) the basic, the random extension and the targeted extension sets (n=1300).

As explained in section 2.2, the exact GHD values are unknown but a set of 100 possible observations was generated for each site. Hence, also the training of the RF model was repeated a hundred times for a specific training set, each time using an alternative set of possible observations. This resulted in 100 alternative RF models. Each of these was then applied to the validation set (n=400, cf. Figure 4) and the set of predicted GHDs was compared with the corresponding set of possible GHD observations to assess model performance. Details of the model evaluation are discussed in section 2.5.

Likewise, each RF model was applied to the GIS layers of considered variables (Table 1), resulting in 100 alternative predicted GHD maps at a 1 km² resolution. From these, we calculated an average GHD map, as well as a map with the total prediction range (maximum – minimum predicted values). The former was expected to provide the best estimate of the GHD for that specific training dataset. The latter provided an indication of the overall uncertainty on the predictions, resulting from uncertainties on the GHD observations. As mentioned in section 2.2, the prediction range map of the RF models trained with only the 'Basic' set was also used to generate the targeted subset (cf. Figure 4).

2.5 Assessing the model performance

Different statistics were used to evaluate the performance of the different RF models, including the Nash-Suthcliffe Model Efficiency (ME; Nash and Sutcliffe, 1970):

$$ME = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - O_{mean})^2} \qquad \text{(Eq. 3)}$$

with n the number of observation sites in the validation dataset, O_i the possible observed GHD of a site, P_i the corresponding predicted GHD and O_{mean} the average of all O_i-values. ME can range between -∞ and 1 and indicates the proportion of observed variance that the model accounts for. A perfect model accounting for all variance has a ME of 1. Negative ME values indicate that the model induces more variance than is initially present in the observations.

While ME is commonly used to evaluate models in geomorphic research (e.g. de Vente et al., 2013; Campforts et al., 2019), it is also highly sensitive to outliers. As complementary measures, we therefore also calculated the overall *Bias* (cf. Eq. 2, section 2.4) as well as the fraction of sites for which the predicted GHD deviated less than five gully heads from its corresponding observed value ('$Frac_5$'). Given that for each of the four different training dataset, 100 alternative RF models were trained (cf. section 2.4), *ME*, *Bias* and $Frac_5$ were calculated for each of these models, using the 100 corresponding sets of possible GHD observations in the validation dataset. For each of these measures, we then calculated the average and 95% range.

To assess the relevance of the considered variables (cf. Table 1), the feature importance was calculated through the '*feature_importances_'* property of the RF regressor (Pedregosa et al., 2011). As discussed in the introduction, RFs consist of a set of decision trees, which are trained by minimizing the impurity of every node through the identification of an optimal decision rule. In a RF regression, the impurity measure is variance. Hence, every decision rule will minimize the variance of samples at the resulting node. Since more relevant features (i.e. potential explanatory variables; cf. Table 1) will tend to be selected earlier in the tree and more often, feature importance quantifies the average reduction in variance per feature over all nodes and trees, weighted by the probability of a sample reaching that node. As such, it indicates the relative importance of a predictor variable within the RF model. Higher values imply that the variable is more important within the model and can therefore be considered as a stronger predictor. Nonetheless, feature importance provides no information about whether

a variable has a positive or negative effect on predicted GHDs. Depending on the node, this effect may also vary.

3. Results

3.1 Observed gully densities and their characteristics

The average observed GHDs across the 1700 observation sites (cf. Figure 4) range between 0 and 525 (Figure 5). While the cumulative frequency distributions are very similar for the basic, validation and random extension set, the targeted extension set is characterized by typically higher GHDs. Figure 5 also indicates the 95% confidence interval of the observed GHDs (calculated as the difference between the 97.5 and the 2.5 quantile of the 100 generated possible GHD values). These intervals span between 2 and 531 heads, but strongly depend on the average density. Observation sites with an average observed GHD of 60 or more heads per km² often have larger intervals than those with a lower GHD, which is generally linked to the occurrence of cells with a score of '3' (cf. section 2.2).

Figure 6 shows the spatial distribution of the average observed GHDs across the Horn of Africa. A majority of sites have a GHD of less than 10 heads km^{-2} (around 70% for the non-targeted subsets; cf. Figure 4, 5). Most of the higher GHD values (>50 or even >100 heads km^{-2}) occur in the Afar region (NE-Ethiopia) and Eritrea.

3.2 Model performance and predicted gully densities

Figure 7 summarizes the key statistics of model performances of the RF models, trained with different datasets. All statistics were calculated using the independent 'validation' set (cf. Figure 4). In terms of Model Efficiency (Figure 7a; cf. Eq. 3), the RF models trained only with the 'basic' set perform poorly. With an average value of 0.02, the variance explained by these models is almost negligible. However, expanding the training set with the 400 sites from the 'random extension' set significantly increases the ME. Adding instead the 400 sites of the 'targeted extension' resulted in an even larger increase in ME. Models trained with all three training datasets (n=1300) clearly showed the highest ME.

In terms of Bias (Figure 7b; cf. Eq. 2), models trained with the basic set + the targeted extension show a slight yet significant bias. For the other training sets, the bias is clearly lower and not significantly different from zero (when considering the 95% range of calculated Bias values). The $frac_5$ values lie around 35% and do not vary much between training sets (Figure 7c). Only models trained with the basic set + targeted extension tend to have a slightly lower $Frac_5$.

Figure 8 shows the average simulated GHDs across the study area at a 1 km² resolution, based on the 100 RF models that were trained with all available training data (i.e. the basic, random extension and targeted extension set; cf. Figure 4). Given that these RF models showed the overall best model performance (cf. Figure 7); this map can be interpreted as our best estimate of GHDs across the Horn of Africa.

Figure 9 shows the difference between the average simulated GHD and the average observed GHD for the 400 validation sites. The background shows the total range in predicted GHDs. For 35% of the sites, average predicted GHD values deviate less than 5 heads from their corresponding average observed GHD. For 56% of the sites this deviation is less than 10 heads, while for 71% of the sites this is less than 20 gully heads. For 11% of the sites, the difference between average observed and predicted values exceeds 50 heads. For around 4%, this difference is larger than 100 gully heads. In agreement with Figure 7b, slightly more sites are overpredicted than underpredicted but both cases occur. The majority of sites showing an important over- or underprediction are located in areas where also the range in predicted values is larger (indicating a larger uncertainty on the predictions).

Figure 10 shows the average feature importance of the considered potential explaining variables (Table 1) as well as their 95% range across the 100 trained RFs. These feature importances vary relatively little in relation to the subset of training data used (cf. Figure 4). Only the RF models trained with the basic set + targeted extension, show a slightly different pattern. Overall, NDVI clearly has the highest feature importance, followed by the annual rainfall depth (P_a). For most other variables, feature importance varies typically between 0.03 and 0.07. Only for CL, the feature importance is close to zero.

In general, feature importances also depend on the overall number of variables considered and their intercorrelation. To gain more insight into the role of the different environmental factors controlling GHD across the Horn of Africa, Figure 11 shows the average feature

importance, stacked according to the overall environmental factor to which they mainly relate (cf. Table 1). This figure indicates that variables describing the topographic and geomorphic context as well as variables describing soil characteristics are most important in explaining spatial variations in GHD (each having a combined feature importance of around 30%). Variables relating to land cover/land use have a combined feature importance of around 22%. Of these variables, NDVI is clearly the most important one. Variables relating to rainfall characteristics have a combined feature importance of about 19%, with variables relating to rainfall intensity (i.e. RDN and P_{99d}) being clearly less important than annual rainfall.

4. Discussion

4.1 Model performance and reliability

Overall, our validation results (Figure 7) show that model performances strongly depend on the amount of data used to train the model. For RF models trained with only 500 observation sites, the variance explained by the model was close to zero. However, models trained with larger datasets yielded significantly higher ME-values (Figure 7a). Differences in Bias (Figure 7b) and Frac5 (Figure 7c) are relatively smaller when different training sets are considered. This indicates that the gains in ME when using larger training datasets are mainly attributable to less severe over- or underestimations in a limited number of sites.

When using the maximum amount of available training data, the trained random forests have a ME of around 20% (Figure 7a). While this may seem low at first sight, it is interesting to put this figure into perspective. First and foremost, there are (to our knowledge) no studies yet that have aimed to predict GD at regional to continental scales and allow for a direct comparison. Several studies have attempted to predict other geomorphic processes at such scales, including sheet and rill erosion and catchment sediment yields. However, they generally provide no direct quantitative validation with an independent validation dataset (e.g. de Vente et al., 2013; Vanmaercke et al., 2014; Borrelli et al., 2017; Fenta et al., 2020). As compared to studies that do provide an independent validation, our model results are similar to (and in several cases even better than) studies aiming to predict spatial patterns of catchment sediment yields (in Ethiopia and elsewhere, cf. de Vente et al., 2013). This is especially so, when taking into account that ME (cf. Eq. 3) was calculated based on absolute rather than log-transformed values.

Nevertheless, several studies did predict gully occurrence over smaller areas and reported statistics on model performances. These results remain difficult to compare as they are based on aggregated gully densities for larger catchments (e.g. Zhao et al., 2016) or only consider the presence or absence of gullies at the pixel level and are based on much smaller validation datasets (e.g. Pourghasemi et al., 2017; Rahmati et al., 2017; Arabameri et al., 2019; Hosseinalizadeh et al., 2019). Nonetheless, they do suggest overall better model performance. To a large extent, this may be attributable to the smaller range in potential factors that may influence gully erosion. In contexts where climatic condition, soil characteristics and/or land use systems vary much less, it might be feasible to predict the occurrence of gullies more precisely; especially since the controlling factors of gully initiation may show important interactions (e.g. Torri & Poesen, 2014; Rossi et al., 2015). For example, when focusing on a specific gullied catchment, it might be possible to predict fairly accurately where gullies may occur, based on the topography and land use conditions (e.g. Rahmati et al., 2017; Hosseinalizadeh et al., 2019). At regional scale, however, areas with a similar topography and land use, may be unaffected by gullies due to their different climatic conditions, soil characteristics or geomorphic/tectonic context. As our results indicate (e.g. Figure 10 & 11), variables that relate to these more regional factors, tend to have a large importance within the trained RFs, making accurate predictions at the level of individual sites potentially harder.

While the performance statistics indicate that our RFs are not very good at accurately predicting the GHD of individual sites (Figure 7), our model is clearly capable of simulating the regional patterns of GHD across the Horn of Africa in a robust and reliable way (with 71 percent of the sites having a predicted average GHD that deviated less than 20 heads from the average observed value; Figure 9). Also a visual comparison between the mapped (Figure 6) and simulated gully density (Figure 8) shows that the model is generally capable of reproducing the main gully density hotspots of the Horn of Africa. However, within these susceptible areas, GHD can vary greatly between individual sites. As Figure 9 indicates, it is mainly in these areas that significant prediction errors occur. Interestingly, also the range in predicted values at pixel level (based on the 100 alternative RF models) is large in these areas (Figure 9). This suggests that our modelling approach not only results in acceptable predictive accuracies, but can also correctly indicate where the uncertainty on predictions is higher. As our validation results based on different subsets demonstrate (Figure 7), this can guide future sampling for further model improvement.

Overall, we deem our map useful for large scale assessments (e.g. comparing gully densities at the catchment scale). Nevertheless, predicted GHD at the pixel/site level can be subject to large uncertainties and should be interpreted with caution.

4.2 Gully densities across the Horn of Africa: a geomorphic interpretation

According to our simulations (Figure 8), potential hotspots of gully erosion mainly occur in Eritrea, the Northern Ethiopian highlands (including Tigray and Amhara), parts of Somali State and the Afar Triangle. To a significant extent, these patterns correspond to predicted patterns of sheet and rill erosion (Fenta et al., 2020). This is to be expected, since the main factors driving sheet & rill erosion (i.e. significant topography, poor vegetation cover, erodible soils and erosive rainfall events; e.g. Fenta et al., 2020; Borrelli et al., 2017) can also lead to gully development (e.g. Torri & Poesen, 2014; Vanmaercke et al., 2016). Nonetheless, there are also significant differences. For example, while our simulations (Figure 8) and mapping efforts (Figure 7) indicate high gully densities in the Afar region, sheet and rill erosion rates are predicted to be relatively low in this area (Fenta et al., 2020). Given the low rainfall amounts in this region (and the associated lack of cultivation), careful interpretation is needed here. Gully heads that formed during an extreme rainfall event are likely preserved in the landscape for many years or even centuries. This may result in a potential observation bias, with the higher GHD being the result of a longer 'representative time period'. A similar issue was reported with respect to the occurrence of landslides and rockfalls (Broeckx et al., 2018; Broeckx et al., 2020), where observed landslides in arid regions are most likely relicts that are no longer attributable to current environmental conditions. Furthermore, while (semi-)arid regions are often characterized by high gully densities (e.g. Poesen et al., 2003; Vanmaercke et al., 2011; Castillo & Gomez, 2016), the highest gully headcut retreat rates are typically observed in more tropical and humid environments (Vanmaercke et al., 2016). As such, a high GHD does not necessarily imply that actual gully erosion rates are high.

Apart from the Afar region, there also appear to be more subtle differences between predicted hotspots of sheet and rill erosion (Fenta et al., 2020) and hotspots of GHD (Figure 8). While this is outside the scope of this paper, it would be interesting to further explore these differences and their relative importance in explaining spatial patterns of catchment sediment yield.

Overall, the nature of the applied RF approach does not allow directly characterizing the importance of different environmental factors. This because explanatory variables may have a different role, depending on the value of other variables. However, herein probably also lies the strength of this method, given that gully erosion is typically a threshold-dependent and non-linear process that can depend on a specific combination of local and regional environmental and geomorphic variables (e.g. Torri & Poesen, 2014; Castillo and Gomez, 2016; Vanmaercke et al., 2016). Nevertheless, the derived feature importance of our RF models does provide some highly interesting and relevant insights (cf. Figure 10 & 11).

In general, variables describing the topographic and geomorphic setting of study sites (Table 1) play an important role in the RF models. Surprisingly, the slope steepness of study sites has a relatively lower importance as compared to other factors. For example, the feature importance of the calculated median k_{sn} values is about 50% higher than for the average or maximum slope steepness (Figure 11). As k_{sn} integrates the effect of both slope steepness and contributing area (cf. Eq. 1), it likely provides a better proxy for the topographic conditions leading to gully initiation (Torri and Poesen, 2014). The relatively high importance of profile curvature provides a further indication for this (Figure 11). This indicates that k_{sn} can provide a geomorphically meaningful and statistically powerful tool to characterize topographic contexts of gully initiation, even in situations where the exact position of gully heads (and, by extent, their specific combination of slope steepness and contributing area) are unknown. At a more fundamental level, the feature importance of k_{sn} suggest that fluvial processes not only control the pace of landscape evolution over geological timescales (e.g. Campforts et al., 2015) but also control the intensity of current hillslope processes. Next to k_{sn}, also the average distance to rivers has a relatively high feature importance (Figure 10; 11). Also this corresponds to our geomorphic understanding that gullies often form in alluvial areas as a result of saturation excess and subsurface flow (e.g. Amare et al., 2019), due to regressive erosion of river knickpoints (e.g. Mendez-Duarte et al., 2007) or due to regressive erosion starting as bank gullies at steep river channel banks (Oostwoud Wijdenes et al., 2000) . It also further demonstrates that, despite similarities, the formation of gullies may be driven by clearly different mechanisms and factors than sheet and rill erosion.

Variables relating to land cover are likewise very significant, with NDVI being the variable with clearly the highest feature importance across all simulations and training sets (Figure 10, 11). This is in line with our general understanding that decreases in vegetation cover can lead to gully formation (e.g. Torri and Poesen, 2014; Zhao et al., 2016), but also points towards

the sensitivity of gully erosion to environmental change. Somewhat surprisingly, the rainfall-weighted version of average NDVI did not perform better. This may be due to errors and uncertainties associated with the proxy or because this proxy does not correctly capture the interactions between rainfall, vegetation cover and erosion. For example, in the semi-arid northern Ethiopian Highlands, the most intense erosion events typically occur at the start of the rainy season when vegetation cover is still low (e.g. Vanmaercke et al., 2010). However, other studies in more humid areas have reported intense alluvial gully erosion at the end of the rainy season, when the vegetation cover is high but soils are also more saturated (e.g. Amare et al., 2019). Similarly, the variable indicating whether a site was predominantly cultivated showed no importance (Figure 10). It was hypothesized that this distinction may be relevant since cropland in the study area is also more frequently treated with soil and water conservation measures (cf. section 2.3; e.g. Hargeweyn et al., 2015), resulting in often relatively lower soil erosion rates as compared to rangeland (e.g. Taye et al., 2015; Maetens et al., 2012a; 2012b). Nevertheless, soil cultivation and especially tillage as well as the creation of drainage ditches on cropland, can also significantly increase the occurrence of gullies (Monsieurs et al., 2016). Most likely, the considered CL variable is too crude as a proxy to accurately reflect such different land management effects.

Also soil characteristics are clearly important in the generated RF models (Figure 10, 11). Nonetheless, especially here, these results are difficult to interpret. Overall, subtle differences in feature importances also depend on the training dataset used as well as inter-correlations with other variables, impeding a full understanding. Moreover, soil characteristics can play different and sometimes contrasting roles in the formation of gullies. They can greatly influence runoff coefficients and by extent the runoff volumes that may be accumulated at a potential gully head. In this regard, high soil bulk densities, high clay and rock fragments and limited soil depths may favour the formation of gullies. On the other hand, the actual formation of gully heads can also strongly depend on the overall soil cohesion and erodibility (Knapen et al., 2007). As such, also deep and less cohesive soils (containing high percentages of sand or silt and having a low rock fragment cover) may be more susceptible to gully erosion. In addition, also the soil characteristics considered here were derived from machine learning approaches and may be subject to important uncertainties (Table 1; Hengl et al., 2017). This further makes the interpretation of these results complex. Here we also used properties estimated for the top soil. Nonetheless, gully erosion can also be controlled by soil properties at larger depths.

Finally, also the variables relating to rainfall characteristics are of significant importance (Figure 10, 11). Annual rainfall appears to be the most relevant variable. Given that both gully head initiation (e.g. Hayas et al., 2017) and headcut retreat (e.g. Vanmaercke et al., 2016) are mainly controlled by rainfall intensity, the relatively lower importance of RDN and P_{99d} is somewhat surprising. As with soil characteristics and other environmental factors, the obtained feature importances also depend on the other variables considered. For example, the feature importance of either RDN or P_{99d} would be higher if only one of the two was considered. Furthermore, the greater importance of annual rainfall may be attributable to its close link to vegetation cover as well as to the potential 'inheritance' effect of gullies in arid environments, discussed above. More research may further constrain and clarify the role of climatic variables. Nonetheless, the dominant importance of NDVI and annual rainfall in explaining GHD (Figure 11) further points to the potentially large sensitivity of gully erosion to climate change in the Horn of Africa (e.g. Vanmaercke et al., 2016).

4.3 Towards gully erosion assessments at larger scales

Despite significant uncertainties at the level of individual sites, this work clearly indicates that it is feasible to simulate gully densities at regional to continental scales with accuracies that are generally comparable to those of other geomorphic processes (e.g. de Vente et al., 2013). While earlier studies already demonstrated the potential of RFs and other machine learning techniques to predict gully occurrence at local scales (e.g. Rahmati et al., 2017; Arabameri et al., 2018; Hosseinalizadeh et al., 2019), their applicability at larger scales remained largely untested. This is likely attributable to the lack of sufficient representative training and testing data. As such, our developed random sampling procedure opens promising perspectives to assess gully densities at regional, continental and even global scales. Here we discuss some scopes for further improvement in this regard.

An important compromise of our semi-quantitative mapping method is that the estimated GHDs are also subject to important uncertainties (Figure 5). These errors propagate in the RF models. Better results might therefore be obtained by further refining the scoring approach, e.g. by increasing the number of cells per site (cf. Figure 2) and/or limiting the possible range of gully heads that correspond to a specific score (cf. Figure 3). Nevertheless, we expect that the potential gain of such refinements will be limited. First, even when gully heads are exactly counted, these quantifications will remain subject to mapping and interpretation

errors (e.g. Maugnard et al., 2014). Second, these uncertainties on the observations are clearly much smaller than the uncertainties on the simulations. For example, for the validation set (Figure 4), correlating two sets of the 100 generated possible GHD values typically resulted in an r^2-value of around 0.85. Correlating these possible observations with their corresponding predicted GHD values typically resulted in an r^2-value of around 0.20 (cf. Figure 7a). As such, the majority of the unexplained variance is attributable to modelling errors rather than observation errors (e.g. Van Rompaey et al., 2001). Over larger areas, the relative importance of observation errors will likely further decrease. This because, across more contrasting environments with larger ranges in GHD, the signal-to-noise ratio of observed gully densities should increase.

Further reducing observation errors by using more accurate assessment procedures would also require significantly more time and, hence, a lower number of sites that could be mapped with the same amount of effort. The latter is important to consider as visually assessing gully densities over larger areas can be highly labour intensive, even when using a semi-quantitative approach as proposed here. On the other hand, our results clearly show that more training data results in better predictions (Figure 7). While practical constraints forced us to limit the training dataset to 1300 sites, it is likely that further increasing the number of sites would further enhance model performance. In this regard, targeting additional mapping efforts in areas where first model results were uncertain clearly resulted in larger gains in model performance (Figure 7a). However, targeting the sampling efforts too much also involves the risks of introducing biases (Figure 7b). We therefore recommend to further explore how the size of the training datasets influences model performances, as well as to further optimize sampling procedures.

Also expanding and fine-tuning the list of potential explanatory variables might lead to further increases in predictive accuracies. We therefore recommend exploring and test other variables that may help characterizing patterns of gully densities at regional to continental scales but that are preferably also straightforward in their interpretation. Examples include variables that: (i) better describe the implementation of soil and water conservation practices as well as other relevant land use practices (e.g. tillage); (ii) better capture the role of vegetation dynamics on gully erosion; (iii) characterize the spatial patterns of vegetation/land cover within study sites (as also this may greatly influence the formation of gullies; e.g. Rossi et al., 2015); (iv) allow to better distinguish the effects of soil properties on runoff

production, erodibility and subsurface flow; and (v) describe the potential effect of tectonics and/or base level changes as well as the overall geology/lithology.

Evidently, increasing the list of potential variables also involves the risk of overfitting or preventing the algorithm of finding an optimal solution. In this study, the potential predictor variables were incorporated without prior analysis and hyperparameters were arbitrarily set to values for which earlier explorations indicated that they would not induce significant biases or overfitting. A third way of improving model performances might therefore lie in further investigating how adapting these and other details of the RF algorithm affects model performances. In this regard, it might also be worthwhile to further explore other machine learning approaches (e.g. Rahmati et al., 2017; Hosseinalizadeh et al., 2019).

5. Conclusions

Accurate simulations and quantifications of gully erosion at regional to continental scales remain a key challenge. This is especially so since gully erosion often depends on a complex combination of interacting environmental factors. Earlier work indicated that machine learning approaches like random forests offer great potential in this regard. However, so far, they have only been tested at local scales. A main constraint for their application at larger scales is large data demands required for training and testing these models.

Here we developed and tested a methodological framework to simulate gully densities across a 1.2 million km² study area in the Horn of Africa, using openly available data and tools and requiring a feasible amount of training data. The main innovations of this approach are: (i) using the number of gully heads rather than gully length or the areal extent as a proxy for gully density; (ii) quantifying these gully heads with a semi-quantitative scoring approach; (iii) sampling gully densities across a large number of small but representative sites, using a combination of random and targeted sampling; and (iv) applying multiple random forests models, trained with alternative sets of observations in order to account for observation uncertainties and their effect on simulated patterns of gully density.

This approach resulted in the first gully density map for Ethiopia, Eritrea and Djibouti (Figure 8). While uncertainties at the level of individual sites can be large, our approach clearly succeeded in robustly simulating regional variations of gully density. Moreover, our approach also allowed identifying areas where prediction errors can be expected to be larger

(Figure 9). Interpretation of the feature importances (Figure 10, 11) further revealed similarities but also important differences with expected patterns of sheet and rill erosion. As such, our results open promising perspectives to better assess the relative importance of gullies to total erosion rates and catchment sediment yields.

In general, we demonstrated that it is feasible to simulate patterns of gully densities at regional, continental or even global scales, using a limited yet representative datasets of mapped gully densities. As discussed in section 4.3, model performances can be expected to further increase by increasing the amount of training data and improving site selection procedures (e.g. through targeted sampling strategies), considering more or better potential explanatory variables and/or fine-tuning the training procedure.

Acknowledgements

Large parts of this article have been written during the COVID-19 lockdown. We therefore want to dedicate this paper to the many victims of this pandemic, as well as to the numerous health workers and scientists that try to keep it at bay. M. Vanmaercke acknowledges the scholarship received from the Japanese Society for the Promotion of Science (JSPS), allowing him to conduct a research stay at the Arid Land Research Center (Tottori University).

Supplementary Materials

Full-resolution GIS version of the GHD map, presented in Figure 8.

Conflicts of interest

None

Data Availability Statement

The map of estimated average gully head densities at a 1 km² resolution (Figure 8) is available in the supplementary materials of this paper. Other source data can be made available by the authors upon reasonable request.

Author contributions

- MVM was responsible for the conceptualisation of the paper, collected a part of the data, conducted the data analysis and drafted the paper.
- YC and SDG were responsible for a large part of the data collection and the GIS analyses and helped with writing and correcting the manuscript.
- WH contributed to conceptualisation of the study and helped developing the Python code used to conduct our analyses. He further contributed to the writing of this paper.
- BC provided the data on river steepness indices, contributed to the conceptualisation of the paper and helped writing the paper.
- NH, AT and JP contributed to the idea that led to this paper, pointed towards relevant literature and helped writing the manuscript.

References

Amare, S., Keesstra, S., van der Ploeg, M., Langendoen, E., Steenhuis, T., & Tilahun, S. (2019). Causes and controlling factors of Valley bottom Gullies. Land, 8(9), 141.

Amatulli, G., Domisch, S., Tuanmu, M. N., Parmentier, B., Ranipeta, A., Malczyk, J., & Jetz, W. (2018). A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Scientific data, 5, 180040.

Arabameri, A., Pradhan, B., Pourghasemi, H. R., Rezaei, K., & Kerle, N. (2018). Spatial modelling of gully erosion using GIS and R programing: A comparison among three data mining algorithms. Applied sciences, 8(8), 1369.

Arabameri, A., Chen, W., Loche, M., Zhao, X., Li, Y., Lombardo, L., ... & Bui, D. T. (2019). Comparison of machine learning models for gully erosion susceptibility mapping. Geoscience Frontiers.

Avni, Y. (2005). Gully incision as a key factor in desertification in an arid environment, the Negev highlands, Israel. Catena, 63(2-3), 185-220.

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I., ... & Adler, R. F. (2019). MSWEP V2 global 3-hourly 0.1 precipitation: methodology and quantitative assessment. Bulletin of the American Meteorological Society, 100(3), 473-500.

Borrelli, P., Robinson, D. A., Fleischer, L. R., Lugato, E., Ballabio, C., Alewell, C., ... & Bagarello, V. (2017). An assessment of the global impact of 21st century land use change on soil erosion. Nature communications, 8(1), 1-13.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Broeckx, J., Vanmaercke, M., Duchateau, R., & Poesen, J. (2018). A data-based landslide susceptibility map of Africa. Earth-Science Reviews, 185, 102-121.

Broeckx, J., Rossi, M., Lijnen, K., Campforts, B., Poesen, J., & Vanmaercke, M. (2020). Landslide mobilization rates: A global analysis and model. Earth-Science Reviews, 201, 102972.

Buchhorn, M., Smets, B., Bertels, L., Lesiv, M., Tsendbazar, N. - E., Herold, M., Fritz, S. (2019) Copernicus Global Land Service: Land Cover 100m: epoch 2015: Globe. Dataset of the global component of the Copernicus Land Monitoring Service 2019. DOI 10.5281/zenodo.3243509

Campforts, B., & Govers, G. (2015). Keeping the edge: A numerical method that avoids knickpoint smearing when solving the stream power law. Journal of Geophysical Research: Earth Surface, 120(7), 1189-1205.

Campforts, B., Vanacker, V., Herman, F., Vanmaercke, M., Schwanghart, W., Tenorio, G. E., Willems, P. and Govers, G. (2019). Parameterization of river incision models requires accounting for environmental heterogeneity: insights from the tropical Andes, Earth Surf. Dyn. Discuss., 2019(September), 1–43, doi:10.5194/esurf-2019-48.

Campo-Bescós, M. A., Flores-Cervantes, J. H., Bras, R. L., Casalí, J., & Giráldez, J. V. (2013). Evaluation of a gully headcut retreat model using multitemporal aerial photographs and digital elevation models. Journal of Geophysical Research: Earth Surface, 118(4), 2159-2173.

Castillo, C., & Gómez, J. A. (2016). A century of gully erosion research: Urgency, complexity and study approaches. Earth-Science Reviews, 160, 300-319.

Chen, W., Pourghasemi, H. R., Kornejady, A., & Zhang, N. (2017). Landslide spatial modeling: Introducing new ensembles of ANN, MaxEnt, and SVM machine learning techniques. Geoderma, 305, 314-327.

Copernicus Service Information (2019) 1km LTS NDVI V2.2 products (1999-2017). Available at: https://land.copernicus.eu/global/products/NDVI

Cox, R., Zentner, D. B., Rakotondrazafy, A. F. M., & Rasoazanamparany, C. F. (2010). Shakedown in Madagascar: Occurrence of lavakas (erosional gullies) associated with seismic activity. Geology, 38(2), 179-182.

De Geeter, S., Vanmaercke, M., Isenborghs, C., & Poesen, J. (2019). Modelling gully erosion rates across Africa: towards a data-driven and process-oriented model. In Geophysical Research Abstracts, Vol. 21, EGU2019-7231-1.

de Vente, J., & Poesen, J. (2005). Predicting soil erosion and sediment yield at the basin scale: scale issues and semi-quantitative models. Earth-science reviews, 71(1-2), 95-125.

de Vente, J., Verduyn, R., Verstraeten, G., Vanmaercke, M., & Poesen, J. (2011). Factors controlling sediment yield at the catchment scale in NW Mediterranean geoecosystems. Journal of Soils and Sediments, 11(4), 690-707.

de Vente, J., Poesen, J., Verstraeten, G., Govers, G., Vanmaercke, M., Van Rompaey, A., ... & Boix-Fayos, C. (2013). Predicting soil erosion and sediment yield at regional scales: where do we stand?. Earth-Science Reviews, 127, 16-29.

de Vente, J., Poesen, J., Verstraeten, G., Van Rompaey, A., & Govers, G. (2008). Spatially distributed modelling of soil erosion and sediment yield at regional scales in Spain. Global and planetary change, 60(3-4), 393-415.

DiBiase, R. A., & Whipple, K. X. (2011). The influence of erosion thresholds and runoff variability on the relationships among topography, climate, and erosion rate. Journal of Geophysical Research: Earth Surface, 116(F4).

Dube, H. B., Mutema, M., Muchaonyerwa, P., Poesen, J., & Chaplot, V. (2020). A global analysis of the morphology of linear erosion features. *Catena*, *190*, 104542.

Fenta, A. A., Tsunekawa, A., Haregeweyn, N., Poesen, J., Tsubo, M., Borrelli, P., ... & Kawai, T. (2020). Land susceptibility to water and wind erosion risks in the East Africa region. *Science of The Total Environment*, *703*, 135016.

Frankl, A., Deckers, J., Moulaert, L., Van Damme, A., Haile, M., Poesen, J., & Nyssen, J. (2016). Integrated solutions for combating gully erosion in areas prone to soil piping: innovations from the drylands of Northern Ethiopia. *Land Degradation & Development*, *27*(8), 1797-1804.

Gayen, A., Pourghasemi, H. R., Saha, S., Keesstra, S., & Bai, S. (2019). Gully erosion susceptibility assessment and management of hazard-prone areas in India using different machine learning algorithms. *Science of the total environment*, *668*, 124-138.

Geurts, P. (2002). Contributions to decision tree induction: bias/variance tradeoff and time series classification (Doctoral dissertation, University of Liège Belgium).

Guyassa, E., Frankl, A., Zenebe, A., Poesen, J., & Nyssen, J. (2018). Gully and soil and water conservation structure densities in semi-arid northern Ethiopia over the last 80 years. *Earth Surface Processes and Landforms*, *43*(9), 1848-1859.

Golosov, V., Yermolaev, O., Rysin, I., Vanmaercke, M., Medvedeva, R., & Zaytseva, M. (2018). Mapping and spatial-temporal assessment of gully density in the Middle Volga region, Russia. *Earth Surface Processes and Landforms*, *43*(13), 2818-2834.

Hammond, M. (2013). The Grand Ethiopian Renaissance Dam and the Blue Nile: implications for transboundary water governance. In *Global Water Forum* (Vol. 1307).

Haregeweyn, N., Poesen, J., Nyssen, J., Verstraeten, G., De Vente, J., Govers, G., ... & Moeyersons, J. (2005). Specific sediment yield in Tigray-Northern Ethiopia: assessment and semi-quantitative modelling. *Geomorphology*, *69*(1-4), 315-331.

Haregeweyn, N., Poesen, J., Nyssen, J., De Wit, J., Haile, M., Govers, G., & Deckers, S. (2006). Reservoirs in Tigray (Northern Ethiopia): characteristics and sediment deposition problems. *Land degradation & development*, *17*(2), 211-230.

Haregeweyn, N., Tsunekawa, A., Nyssen, J., Poesen, J., Tsubo, M., Tsegaye Meshesha, D., ... & Tegegne, F. (2015). Soil erosion and conservation in Ethiopia: a review. *Progress in Physical Geography*, *39*(6), 750-774.

Haregeweyn, N., Tsunekawa, A., Poesen, J., Tsubo, M., Meshesha, D. T., Fenta, A. A., ... & Adgo, E. (2017). Comprehensive assessment of soil erosion risk for better land use planning in river basins: Case study of the Upper Blue Nile River. *Science of the Total Environment*, *574*, 95-108.

Hayas, A., Poesen, J., & Vanwalleghem, T. (2017a). Rainfall and vegetation effects on temporal variation of topographic thresholds for gully initiation in Mediterranean cropland and olive groves. Land Degradation & Development, 28(8), 2540-2552.

Hayas, A., Vanwalleghem, T., Laguna, A., Peña Acevedo, A., & Giráldez, J. V. (2017b). Reconstructing long-term gully dynamics in Mediterranean agricultural areas.

Hengl, T., de Jesus, J. M., Heuvelink, G. B., Gonzalez, M. R., Kilibarda, M., Blagotić, A., ... & Guevara, M. A. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS one*, *12*(2).

Hosseinalizadeh, M., Kariminejad, N., Chen, W., Pourghasemi, H. R., Alinejad, M., Behbahani, A. M., & Tiefenbacher, J. P. (2019). Gully headcut susceptibility modeling using functional trees, naïve Bayes tree, and random forest models. Geoderma, 342, 1-11.

Huffman, G.J., Stocker, EF., Bolvin, D.T., Nelkin, E.J., Tan, J. (2019). GPM IMERG Final Precipitation L3 1 month 0.1 degree x 0.1 degree V06, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: [2020/3/12], 10.5067/GPM/IMERG/3B-MONTH/06

Ionita, I. (2006). Gully development in the Moldavian Plateau of Romania. *Catena*, *68*(2-3), 133-140.

Jurchescu, M., & Grecu, F. (2015). Modelling the occurrence of gullies at two spatial scales in the Olteţ Drainage Basin (Romania). *Natural Hazards*, *79*(1), 255-289.

Kariminejad, N., Hosseinalizadeh, M., Pourghasemi, H. R., Bernatek-Jakiel, A., Campetella, G., & Ownegh, M. (2019). Evaluation of factors affecting gully headcut location using summary statistics and the maximum entropy model: Golestan Province, NE Iran. Science of the Total Environment, 677, 281-298.

Karydas, C., & Panagos, P. (2020). Towards an Assessment of the Ephemeral Gully Erosion Potential in Greece Using Google Earth. *Water*, *12*(2), 603.

Knapen, A., Poesen, J., Govers, G., Gyssels, G., & Nachtergaele, J. (2007). Resistance of soils to concentrated flow erosion: A review. Earth-Science Reviews, 80(1-2), 75-109.

Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, *27*(15), 2171-2186.

Lemma, H., Admasu, T., Dessie, M., Fentie, D., Deckers, J., Frankl, A., ... & Nyssen, J. (2018). Revisiting lake sediment budgets: How the calculation of lake lifetime is strongly data and method dependent. Earth Surface Processes and Landforms, 43(3), 593-607.

Li, Z., & Fang, H. (2016). Impacts of climate change on water erosion: A review. *Earth-Science Reviews*, *163*, 94-117.

Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*.

Maetens, W., Vanmaercke, M., Poesen, J., Jankauskas, B., Jankauskiene, G., & Ionita, I. (2012a). Effects of land use on annual runoff and soil loss in Europe and the Mediterranean: A meta-analysis of plot data. *Progress in Physical Geography*, *36*(5), 599-653.

Maetens, W., Poesen, J., & Vanmaercke, M. (2012b). How effective are soil conservation techniques in reducing plot runoff and soil loss in Europe and the Mediterranean?. *Earth-Science Reviews*, *115*(1-2), 21-36.

Mararakanye, N., & Le Roux, J. J. (2012). Gully location mapping at a national scale for South Africa. *South African Geographical Journal*, *94*(2), 208-218.

Martineli Costa, F. M., & Bacellar, L. D. A. P. (2007). Analysis of the influence of gully erosion in the flow pattern of catchment streams, Southeastern Brazil. *Catena*, *69*(3), 230-238.

Maugnard, A., Van Dyck, S., & Bielders, C. L. (2014). Assessing the regional and temporal variability of the topographic threshold for ephemeral gully initiation using quantile regression in Wallonia (Belgium). *Geomorphology*, *206*, 165-177.

Menéndez-Duarte, R., Marquínez, J., Fernández-Menéndez, S., & Santos, R. (2007). Incised channels and gully erosion in Northern Iberian Peninsula: Controls and geomorphic setting. *Catena*, *71*(2), 267-278.

Merritt, W. S., Letcher, R. A., & Jakeman, A. J. (2003). A review of erosion and sediment transport models. *Environmental Modelling & Software*, *18*(8-9), 761-799.

Monsieurs, E., Poesen, J., Dessie, M., Adgo, E., Verhoest, N. E., Deckers, J., & Nyssen, J. (2015). Effects of drainage ditches and stone bunds on topographical thresholds for gully head development in North Ethiopia. Geomorphology, 234, 193-203.

Monsieurs, E., Dessie, M., Verhoest, N. E., Poesen, J., Adgo, E., Deckers, J., & Nyssen, J. (2016). Impact of draining hilly lands on runoff and on-site erosion: a case study from humid Ethiopia. Earth Surface Processes and Landforms, 41(4), 513-525.

Nachtergaele, J., Poesen, J., Sidorchuk, A., & Torri, D. (2002). Prediction of concentrated flow width in ephemeral gully channels. *Hydrological Processes*, *16*(10), 1935-1953.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of hydrology*, *10*(3), 282-290.

New, M., Lister, D., Hulme, M., & Makin, I. (2002). A high-resolution data set of surface climate over global land areas. *Climate research*, *21*(1), 1-25.

Nyssen, J., Poesen, J., Moeyersons, J., Deckers, J., Haile, M., & Lang, A. (2004). Human impact on the environment in the Ethiopian and Eritrean highlands—a state of the art. *Earth-science reviews*, *64*(3-4), 273-320.

Nyssen, J., Vandenreyken, H., Poesen, J., Moeyersons, J., Deckers, J., Haile, M., ... & Govers, G. (2005). Rainfall erosivity and variability in the Northern Ethiopian Highlands. *Journal of Hydrology*, *311*(1-4), 172-187.

Oostwoud Wijdenes, D., Poesen, J., Vandekerckhove, L., & Ghesquiere, M. (2000). Spatial distribution of gully head activity and sediment supply along an ephemeral channel in a Mediterranean environment. Catena, 39(3), 147-167.

Owens, P. N., Batalla, R. J., Collins, A. J., Gomez, B., Hicks, D. M., Horowitz, A. J., ... & Petticrew, E. L. (2005). Fine-grained sediment in river systems: environmental significance and management issues. *River research and applications*, *21*(7), 693-717.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Pelletier, J. D., Brad Murray, A., Pierce, J. L., Bierman, P. R., Breshears, D. D., Crosby, B. T., ... & Lancaster, N. (2015). Forecasting the response of Earth's surface to future climatic and land use changes: A review of methods and research needs. Earth's Future, 3(7), 220-251.

Poesen, J., Nachtergaele, J., Verstraeten, G., & Valentin, C. (2003). Gully erosion and environmental change: importance and research needs. *Catena*, *50*(2-4), 91-133.

Poesen J, Torri D, Vanwalleghem T. (2011). Gully erosion: procedures to adopt when modelling soil erosion in landscapes affected by gullying. In Handbook of Erosion Modelling, Morgan R, Nearing M (eds). Blackwell, Oxford, 360–386.

Poesen, J. (2018). Soil erosion in the Anthropocene: Research needs. *Earth Surface Processes and Landforms*, *43*(1), 64-84.

Pourghasemi, H. R., Yousefi, S., Kornejady, A., & Cerdà, A. (2017). Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling. Science of the Total Environment, 609, 764-775.

Radoane, M., Ichim, I., & Radoane, N. (1995). Gully distribution and development in Moldavia, Romania. *Catena*, *24*(2), 127-146.

Rahmati, O., Tahmasebipour, N., Haghizadeh, A., Pourghasemi, H. R., & Feizizadeh, B. (2017). Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. Geomorphology, 298, 118-137.

Rosa, L., Chiarelli, D. D., Rulli, M. C., Dell'Angelo, J., & D'Odorico, P. (2020). Global agricultural economic water scarcity. *Science Advances*, *6*(18), eaaz6031.

Rossi, M., Torri, D., & Santi, E. (2015). Bias in topographic thresholds for gully heads. *Natural Hazards*, *79*(1), 51-69.

Settles, B. (2009). Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences.

Sidorchuk, A. (2020). The Potential of Gully Erosion on the Yamal Peninsula, West Siberia. Sustainability, 12(1), 260.

Thompson, J. R. (1964). Quantitative effect of watershed variables on rate of gully-head advancement. *Transactions of the ASAE*, *7*(1), 54-0055.

Torri, D., & Poesen, J. (2014). A review of topographic threshold conditions for gully head development in different environments. *Earth-Science Reviews*, *130*, 73-85.

Torri, D., Poesen, J., Rossi, M., Amici, V., Spennacchi, D., & Cremer, C. (2018). Gully head modelling: A Mediterranean badland case study. Earth Surface Processes and Landforms, 43(12), 2547-2561.

Valentin, C., Poesen, J., & Li, Y. (2005). Gully erosion: impacts, factors and control. *Catena*, *63*(2-3), 132-153.

Vanmaercke, M., Poesen, J., Maetens, W., de Vente, J., & Verstraeten, G. (2011). Sediment yield as a desertification risk indicator. *Science of the Total Environment*, *409*(9), 1715-1725.

Vanmaercke, M., Maetens, W., Poesen, J., Jankauskas, B., Jankauskiene, G., Verstraeten, G., & de Vente, J. (2012). A comparison of measured catchment sediment yields with measured and predicted hillslope erosion rates in Europe. *Journal of Soils and Sediments*, *12*(4), 586-602.

Vanmaercke, M., Poesen, J., Broeckx, J., & Nyssen, J. (2014). Sediment yield in Africa. *Earth-Science Reviews*, *136*, 350-368.

Vanmaercke, M., Poesen, J., Van Mele, B., Demuzere, M., Bruynseels, A., Golosov, V., ... & Fuseina, Y. (2016). How fast do gully headcuts retreat?. *Earth-Science Reviews*, *154*, 336-355.

Vanmaercke, M. Panos, P., Vanwalleghem, T., Hayas, A., Foerster, S., Borrelli, P., … & Poesen, J. (subm.) Measuring, modelling and managing gully erosion at regional to continental scales: a state of the art. *Earth-Science Reviews, submitted.*

Vannoppen, W., Vanmaercke, M., De Baets, S., & Poesen, J. (2015). A review of the mechanical effects of plant roots on concentrated flow erosion rates. *Earth-Science Reviews*, *150*, 666-678.

Wobus, C., Whipple, K. X., Kirby, E., Snyder, N., Johnson, J., Spyropolou, K., ... & Willett, S. D. (2006). Tectonics from topography: Procedures, promise, and pitfalls. Special papers-geological society of america, 398, 55.

Yibeltal, M., Tsunekawa, A., Haregeweyn, N., Adgo, E., Meshesha, D. T., Aklog, D., ... & Ebabu, K. (2019). Analysis of long-term gully dynamics in different agro-ecology settings. *Catena*, *179*, 160-174.

Youssef, A. M., Pourghasemi, H. R., Pourtaghi, Z. S., & Al-Katheeri, M. M. (2016). Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides*, *13*(5), 839-856.

Zhao, J., Vanmaercke, M., Chen, L., & Govers, G. (2016). Vegetation cover and topography rather than human disturbance control gully density and sediment production on the Chinese Loess Plateau. *Geomorphology*, *274*, 92-105.

Zenebe, A., Vanmaercke, M., Poesen, J., Verstraeten, G., Haregeweyn, N., Haile, M., ... & Nyssen, J. (2013). Spatial and temporal variability of river flows in the degraded semi-arid tropical mountains of northern Ethiopia. *Zeitschrift für Geomorphologie*, *57*(2), 143-169.

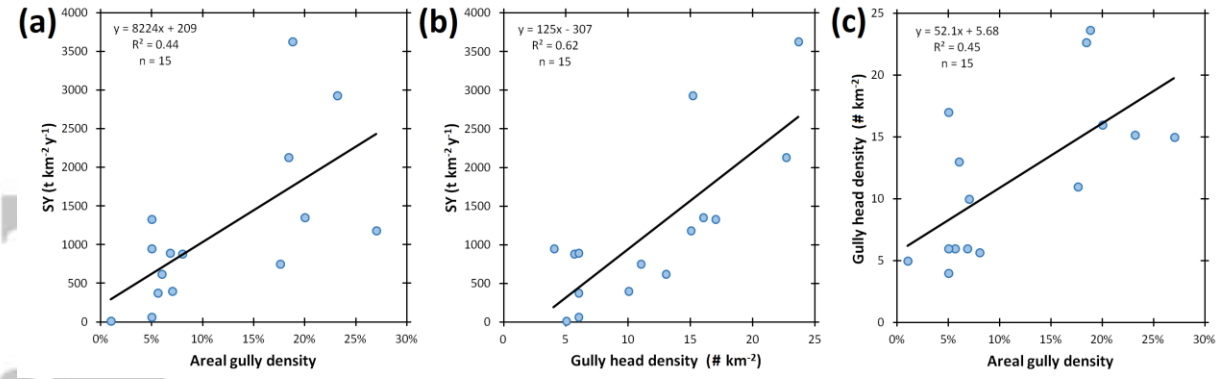


Figure 1*: Comparison of observed catchment sediment yields (SY) and different proxies of gully density for 15 catchments in Ethiopia. SY data were derived from Vanmaercke et al. (2014).* ***(a)*** *SY versus areal gully density.* ***(b)*** *SY versus gully head density.* ***(c)*** *Gully head density versus areal gully density.*

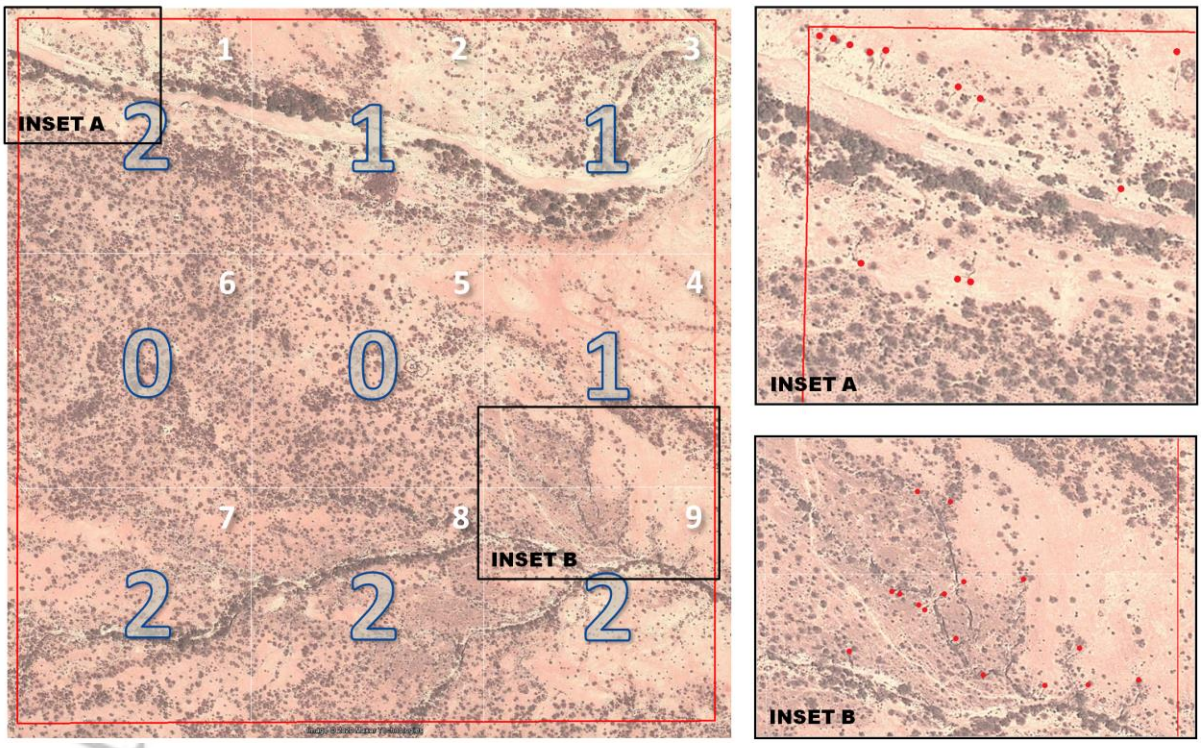


***Figure 2**: Example of a randomly selected observation site in Ethiopia (7.028368°N, 41.94738°E, image date: 05/02/2012). The red square indicates the boundaries of the 1-km² site. The white lines demarcate the boundaries of the nine cells that subdivided each site. The white numbers in the top right corners show the order in which the gully head density of each cell was assessed. The large numbers in the center of each cell indicate the score that was assigned to each cell (where: 0 = no gully heads, 1 = 1-5 gully heads, 2 = 6-20 gully heads, and 3 = >20 gully heads). Insets A and B provide some close-ups of the observation site with visible gully heads marked with a red dot.*

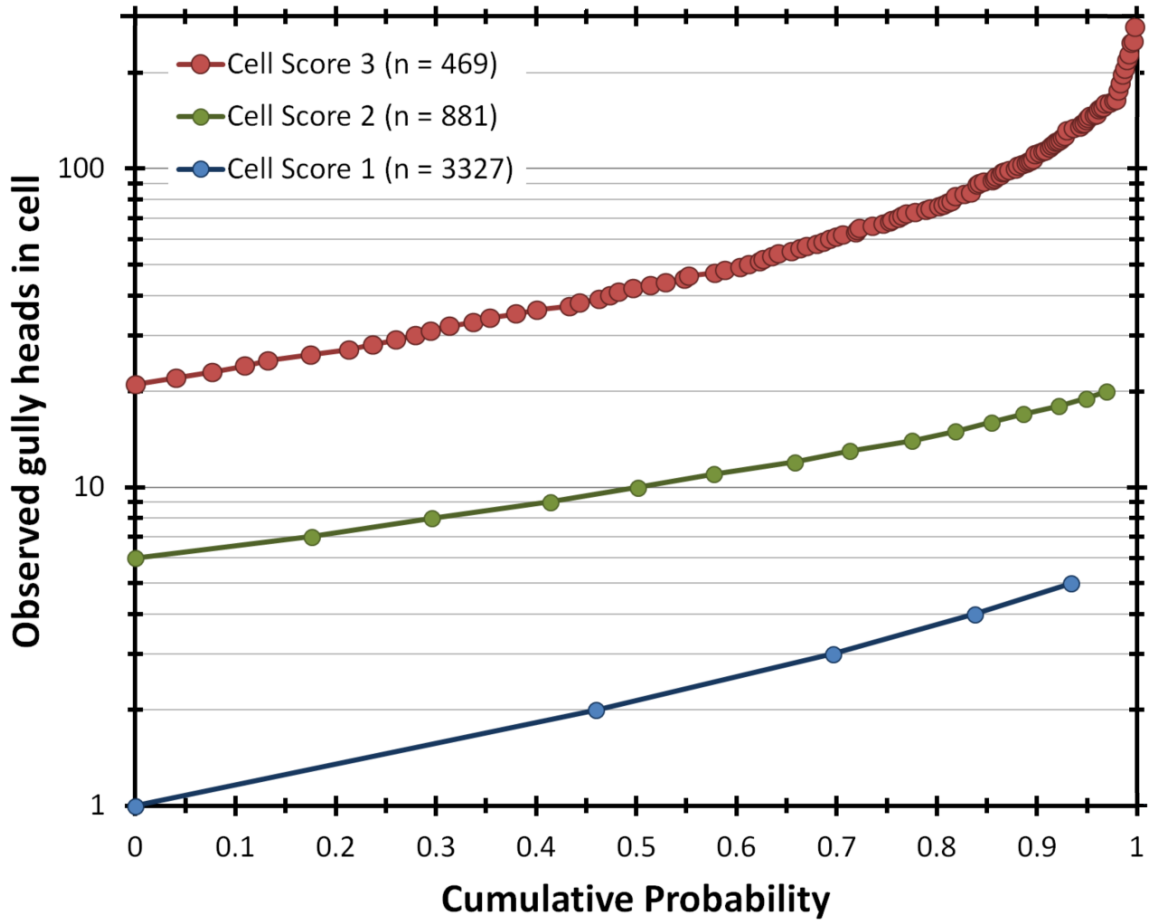


***Figure 3**: Cumulative probabilities of the number of gully heads that can be expected in a cell with the indicated score. These distributions were derived from an earlier developed database of 1 km² study sites across Africa and Europe for which all gully heads were manually mapped in Google Earth (e.g. De Geeter et al., 2019). 'n' indicates the number of cells that were used to calculate the cumulative distributions.*

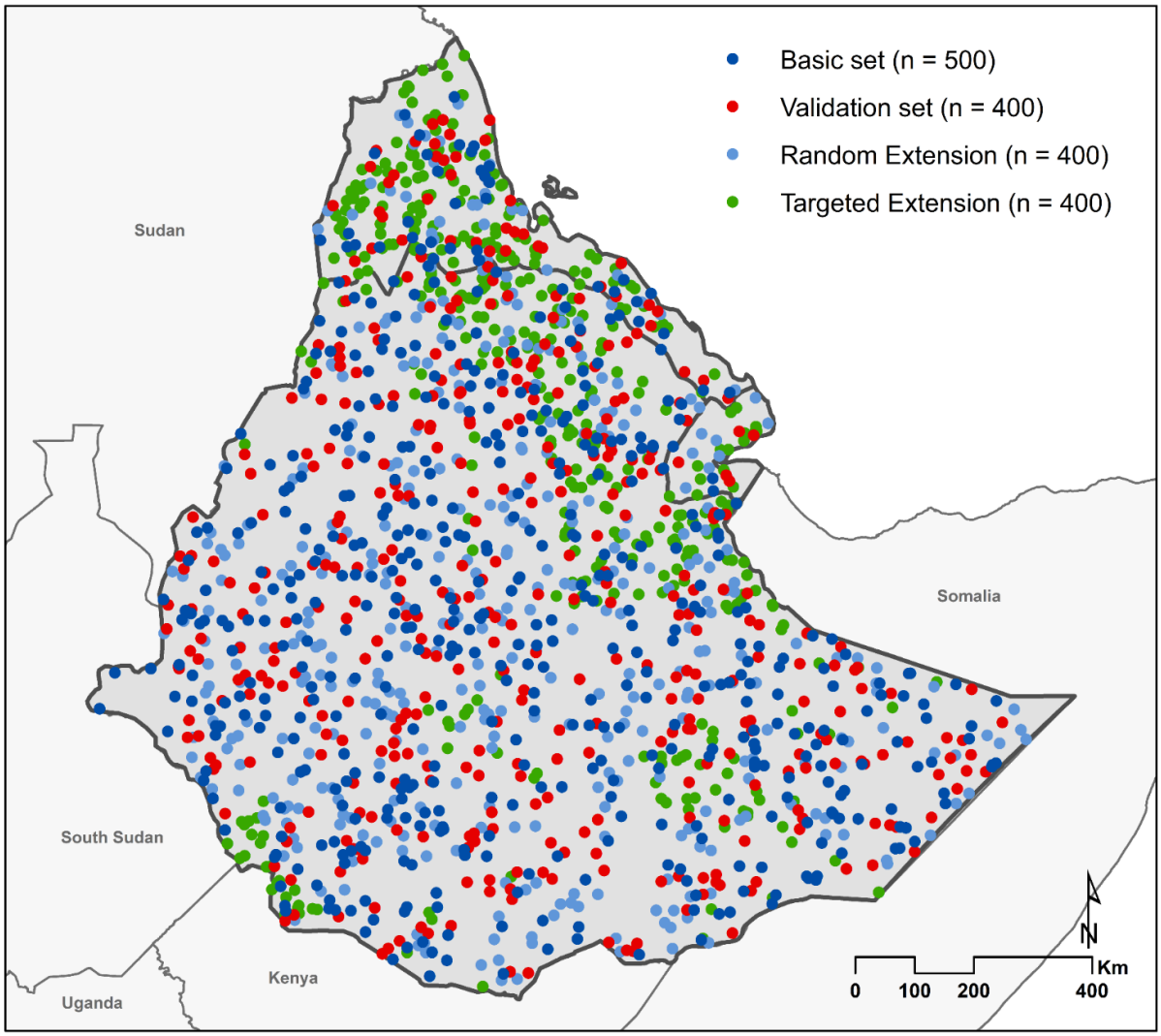


***Figure 4**: Overview of observation sites in the Horn of Africa with possible observations on the gully head density available. Each dot corresponds to a 1 km² observation site (total n = 1700). These sites were subdivided into different subsets for model training and validation (see text for details).*

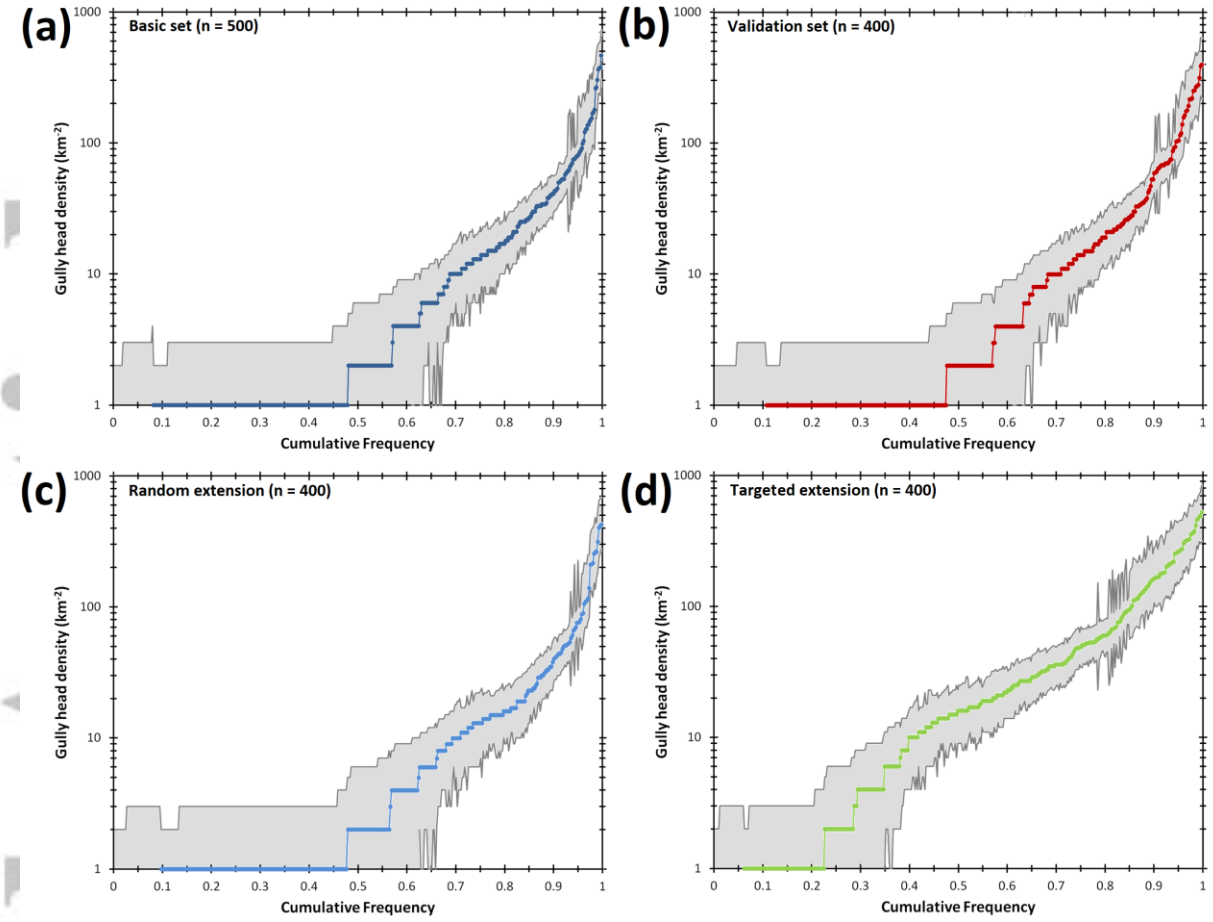


Figure 5: *Cumulative frequency distributions of the observed gully head density (GHD) across different sites for the different datasets indicated in Figure 4. Coloured curves show the average observed GHD (cf. section 2.2). The grey areas indicate the 95% confidence limit on these observations (calculated as the difference between the 97.5 and 2.5 quantile of the possible observed GHD values).*

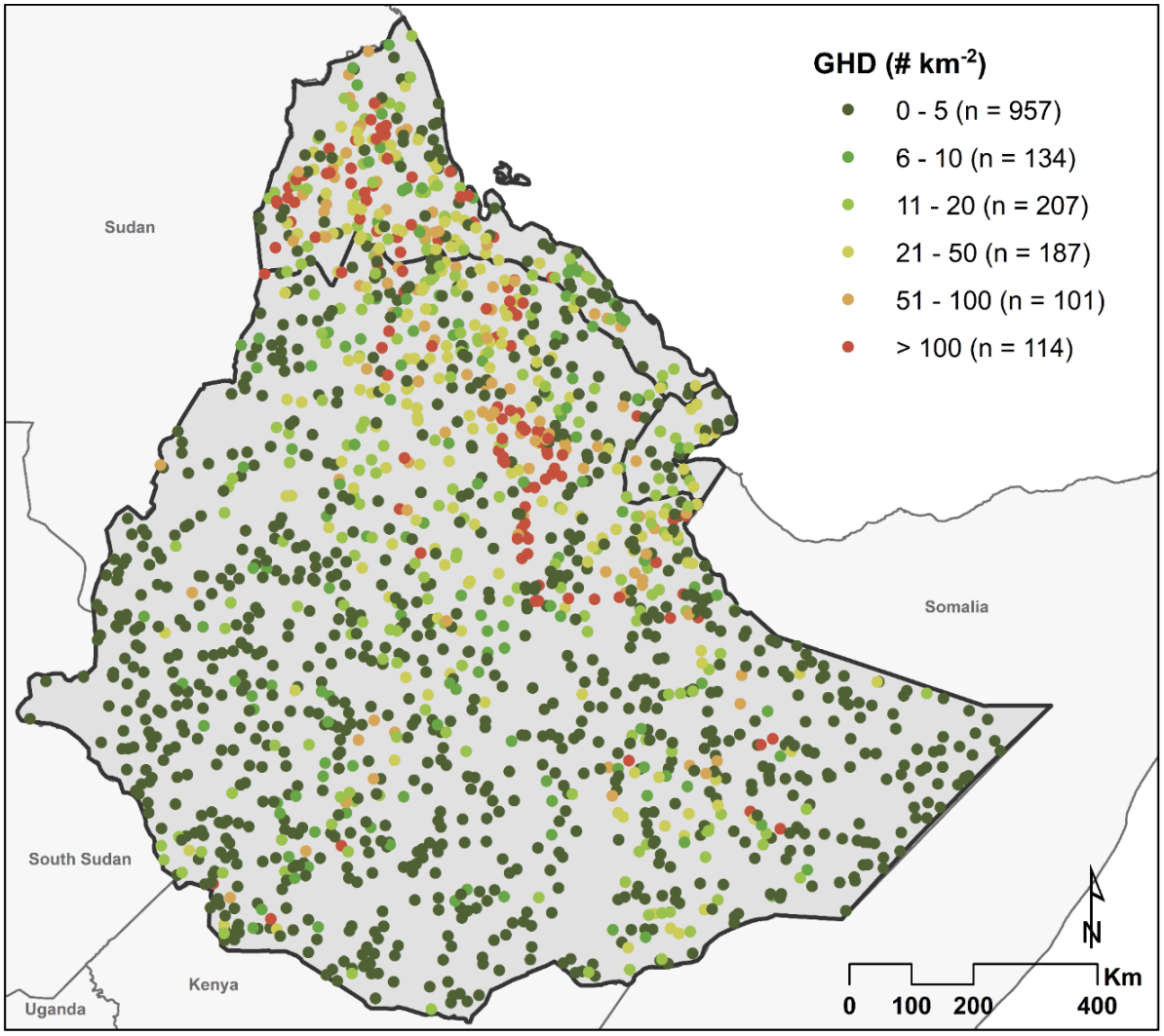


Figure 6: *Spatial distribution of the average observed gully head densities (GHD) of all observation sites (cf. section 2.2).*

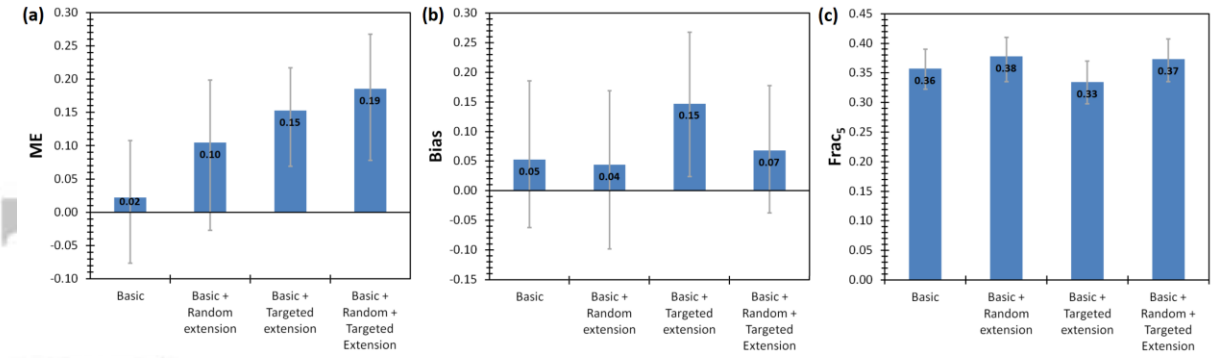


Figure 7: *Model performance statistics for random forest models trained with different subsets of data, applied to the independent validation dataset (cf. Figure 4). Bars indicate the average performance for 100 random forest models, trained with 100 different sets of possible observations. Error bars indicate the 95% range. (a) Nash-Sutcliffe Model Efficiency (cf. Eq. 3). (b) Estimated bias in the total number of gully heads (cf. Eq. 2). (c) Fraction of sites for which the simulated gully head density deviates less than five heads from the corresponding possible observed density.*

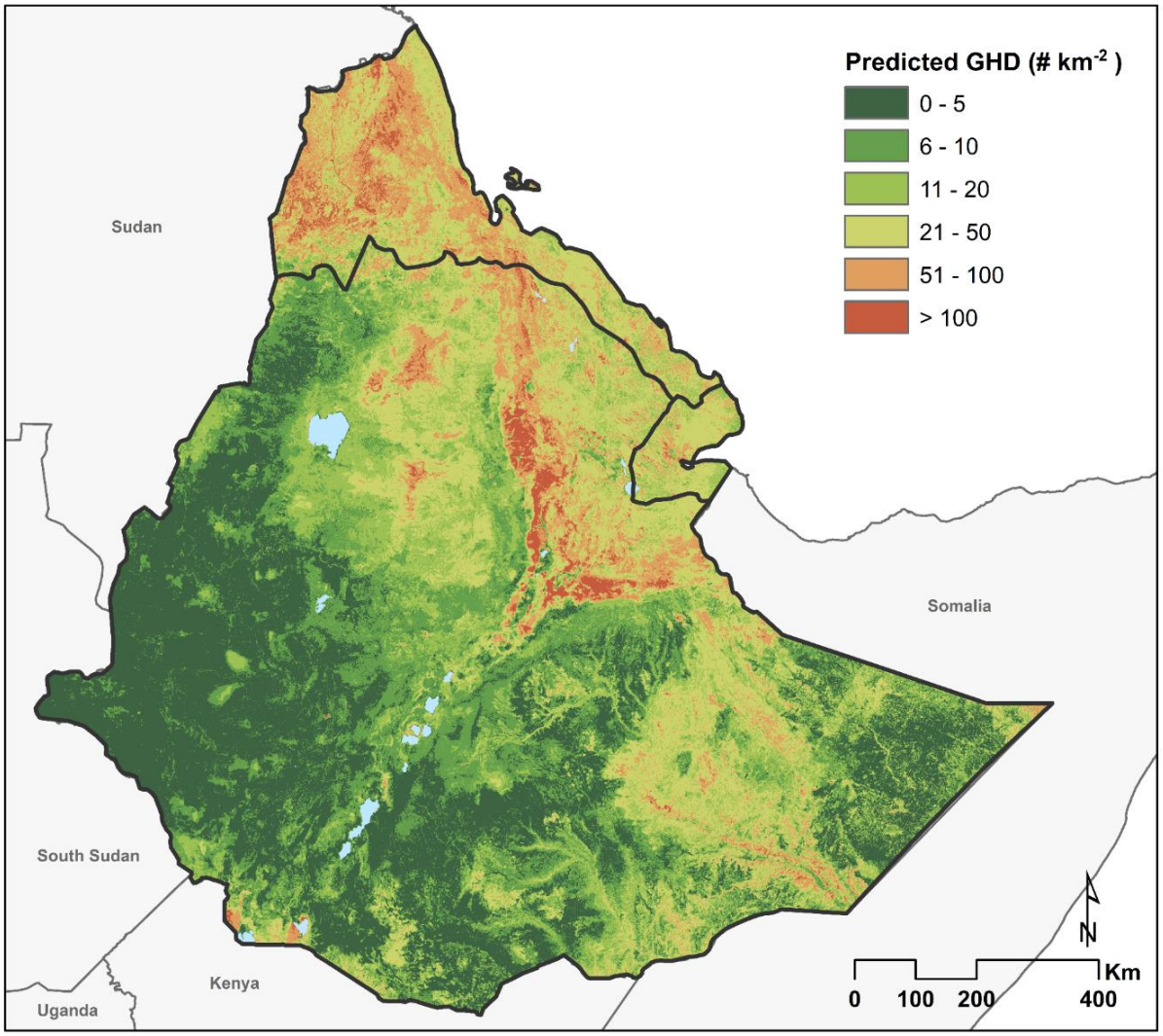


Figure 8: *Simulated gully head densities (GHD) for Ethiopia, Eritrea and Djibouti based on the average of 100 random forest models (trained with the basic, random extension and targeted extension subsets of observation sites; cf. Figure 4).*

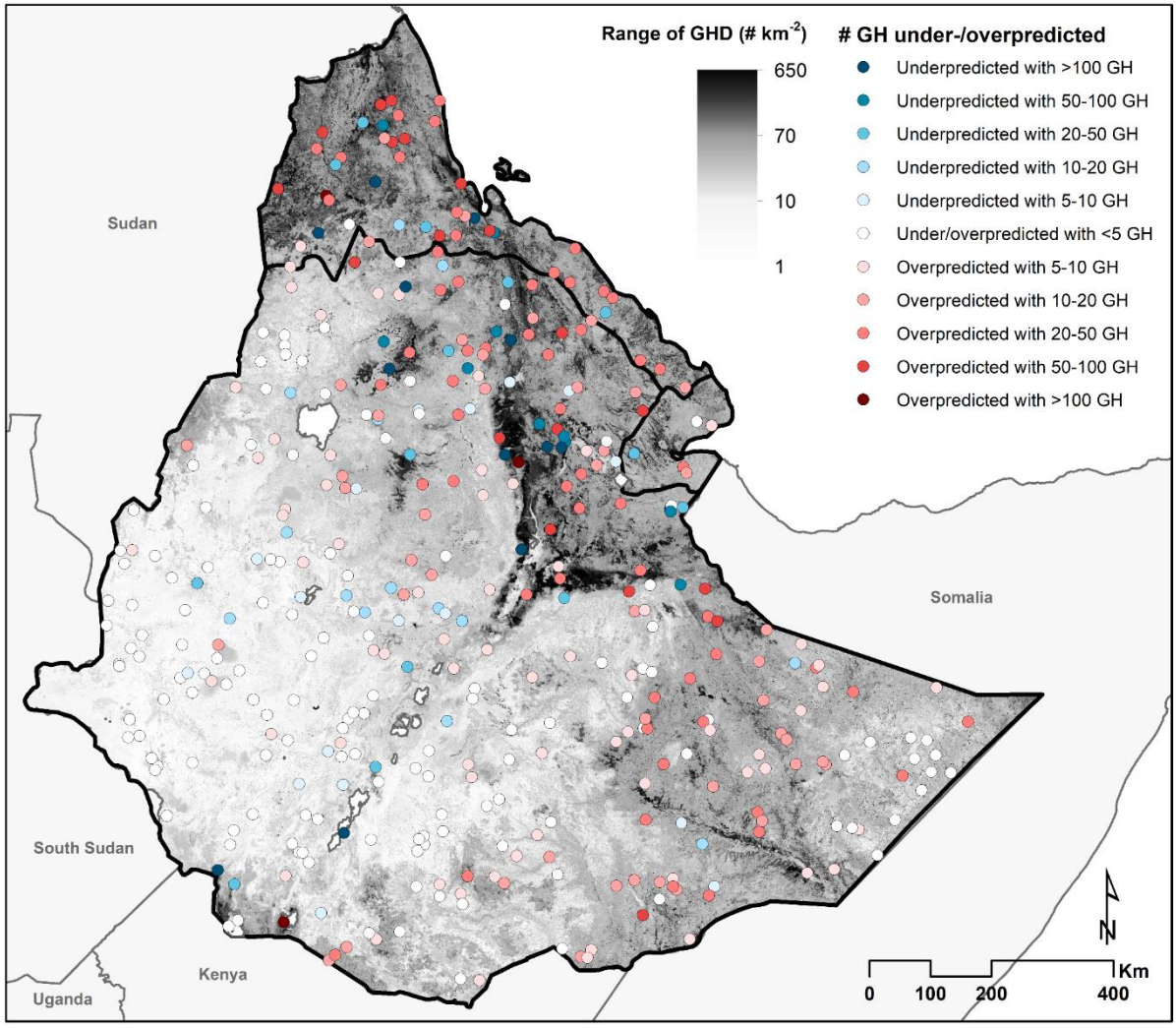


Figure 9: *Difference between the average simulated and average observed gully head density (GHD) for the 400 observation sites of the validation set (cf. Figure 4), based on the same 100 random forest models used to generate Figure 8. For 71% of the sites, this deviation is less than 20 gully heads (GH). For 11%, the deviation is larger than 50 gully heads. The background layer indicates the total range in predicted GHD values, a proxy for predictive uncertainty.*

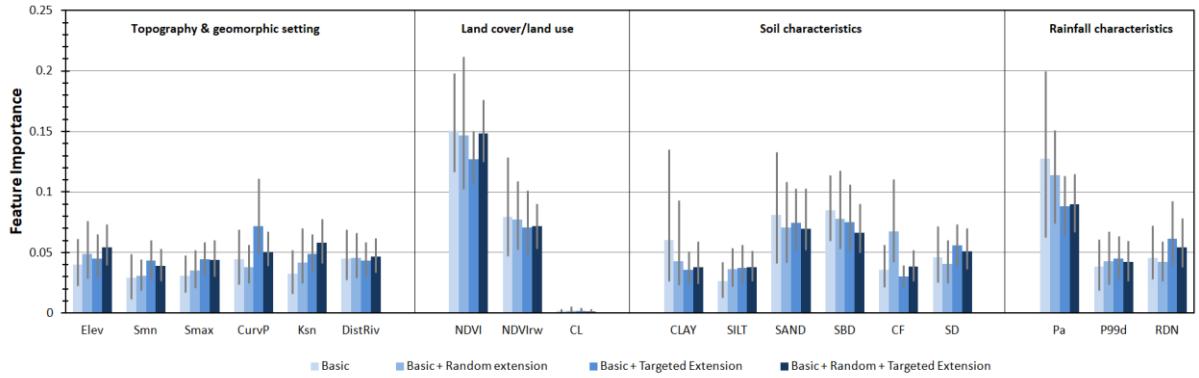


Figure 10: *Feature importance of the considered explanatory variables (see Table 1 for details and an explanation of the abbreviations), grouped according to the four subsets of sites used for training the model (cf. Figure 4). Bar heights indicate the average feature importance, based on the 100 trained random forest models. Error bars indicate the corresponding 95% variability range.*

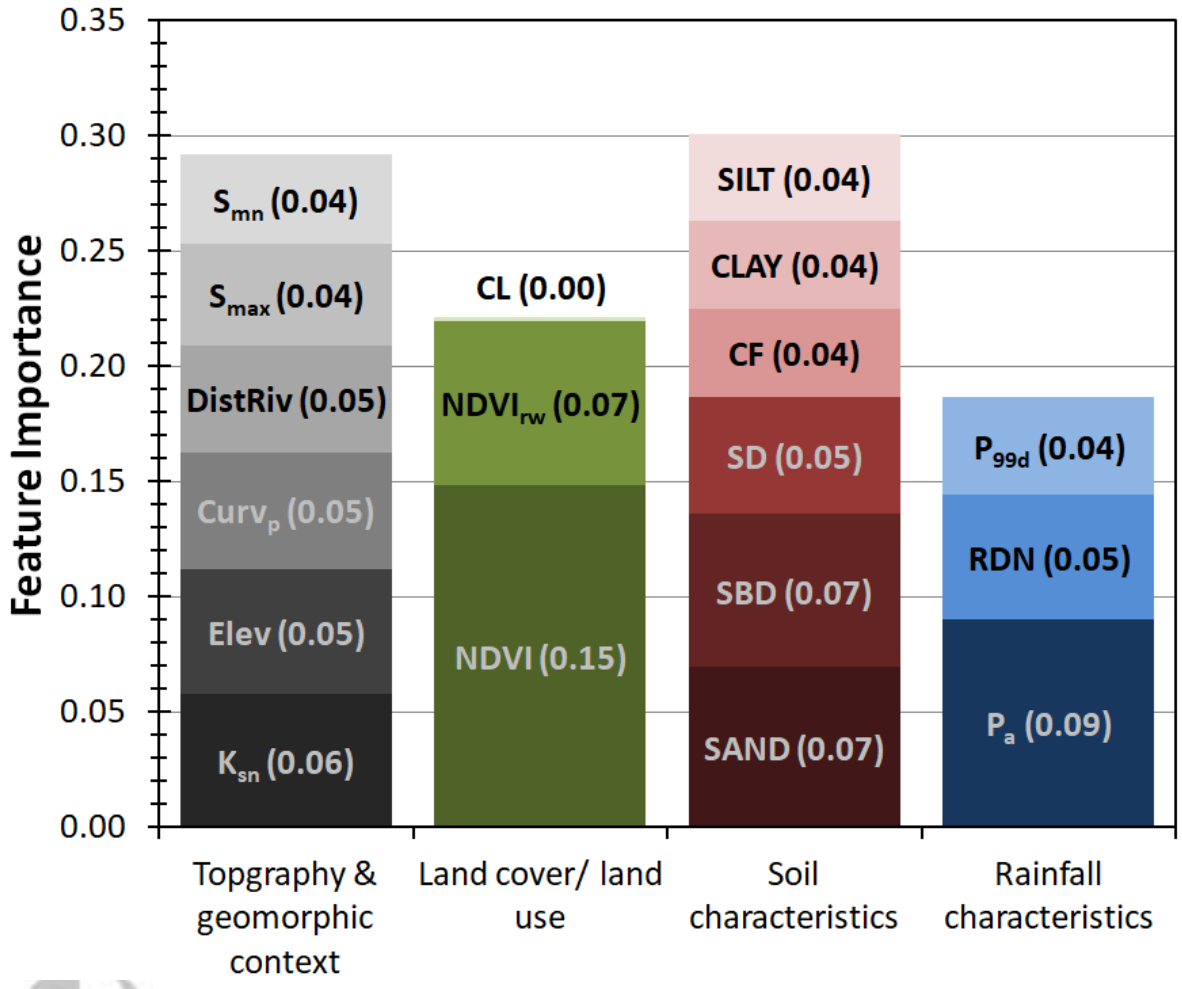
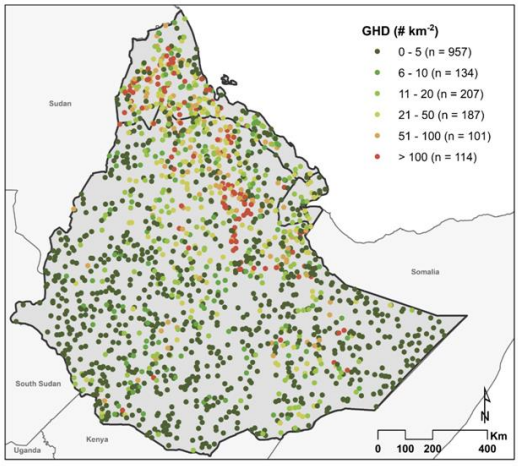


Figure 11: *Average feature importance of the considered explanatory variables (see Table 1 for details and an explanation of the abbreviations), for the random forest models trained with all available training data (i.e. basic set + random extension + targeted extension; cf. Figure 4). Variables are stacked according to the overall environmental factor they characterize (cf. Table 1).*

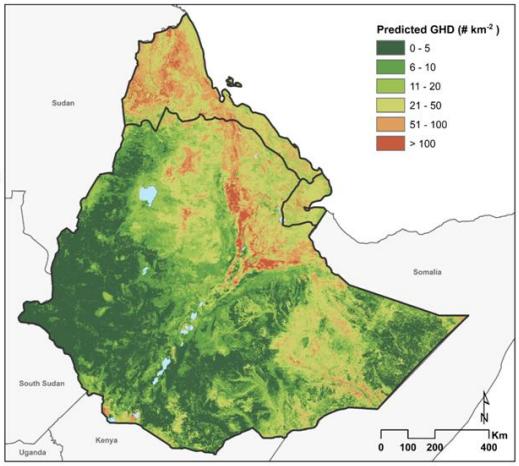
***Table 1**: Overview of the considered variables potentially explaining the variability in estimated gully head densities between observation sites. The column 'overall range' indicates the minimum and maximum value of the variable across all observation sites (n = 1700, cf. Figure 4).*

Factor	Variable	Description	Units	Original resolution	Overall Range	Source
Topography & geomorphic context	Elev	Average elevation of the observation site	m	1000m	-121 - 4291	Amatulli et al. (2018)
	S_{mn}	Mean slope steepness of the observation site	°	1000m	0.29 - 35.15	Amatulli et al. (2018)
	S_{max}	Maximum slope steepness of the observation site	°	1000m	0.70 - 59.35	Amatulli et al. (2018)
	$Curv_P$	Average profile curvature of the observation site	rad m^{-1}	1000m	-0.00094 - 0.00104	Amatulli et al. (2018)
	k_{sn}	Median normalized steepness index of the observation site (cf. Eq. 1)	$m^{0.9}$	90m	0.45 - 97.95	Lehner et al. (2013); own processing
	DistRiv	Average distance of the observation site to a stream with Strahler order ≥ 4	m	500m	171 - 67532	Lehner et al. (2013); own processing
Land cover/land use	NDVI	Long-term (1999-2017) average NDVI of the observation site	none	1000m	0.009 - 0.936	Copernicus Service Information (2019)
	$NDVI_{rw}$	Average NDVI (1999-2017) of the observation sites, weighted according to monthly rainfall	none	1000m	0.009 - 0.936	Copernicus Service Information (2019); Huffman et al. (2019); own processing
	CL	Boolean variable, indicating if the observation site is dominated by cropland or not	none	100m	0 - 1	Buchhorn et al. (2019)
Soil characteristics	CLAY	Average mass percentage of clay in the topsoil	%	250m	3.9 - 100	Hengl et al. (2017)
	SILT	Average mass percentage of clay in the topsoil	%	250m	3.9 - 100	Hengl et al. (2017)

	SAND	Average mass percentage of sand in the topsoil	%	250m	6.7 - 100	Hengl et al. (2017)
	SBD	Average soil bulk density	kg/m^3	250m	880 - 1549	Hengl et al. (2017)
	CF	Average volumetric coarse fragment content	%	250m	0 - 100	Hengl et al. (2017)
	SD	Average soil depth	cm	250m	505 - 18612	Hengl et al. (2017)
Rainfall characteristics	P_a	Average annual rainfall (1979-2016)	mm y^{-1}	0.1°	42 - 1844	Beck et al. (2019); Broeckx et al. (2020)
	P_{99d}	99% percentile of the daily precipitation in the period 1979-2016	mm day^{-1}	0.1°	2 - 53	Beck et al. (2019); Broeckx et al. (2020)
	RDN	Rainy day normal, i.e. average annual precipitation divided by the average number of rainy days (for the period 1961-1990)	mm rainy day^{-1}	10'	4.22 - 19.60	New et al. (2002); Vanmaercke et al. (2016)



Random Forest regression approach



- Novel and fast way to map gully head density (GHD)
- Using Google Eath and a visual scoring approach
- GHD mapped for 1,700 sites

- First gully density map for the Horn of Africa
- Acceptable model performance
- Viable strategy to assess gully erosion risks at continental scales