Supplementary Materials: Identifying Soil Erosion Processes in Alpine Grasslands on Aerial Imagery with a U-Net Convolutional Neural Network

Maxim Samarin†, Lauren Zweifel†, Volker Roth and Christine Alewell

S1. Details on the Neural Network Architecture

The main components of the U-Net architecture are convolution, max pooling, dropout, and transposed convolution operations with rectified linear unit (ReLU) activations. In the following, we give some more details on the individual components of the neural network and how these components are combined in the U-Net architecture.

The convolution operation is represented by a kernel (typically of size 3×3) which processes the input image leading to an intermediate representation of activations called the feature map (illustrated in Figure S1 (b)). Using several kernels enables identification of different meaningful features in the input images which are learned during the training of the neural network. Pooling operations combine adjacent pixels into a summary statistic and sub-sample feature maps, a process which effectively reduces the size of the feature map. For instance, the max pooling operation with a 2×2 kernel takes the activation of four adjacent pixels in the feature map and stores only the maximum value in the subsequent feature map (depicted in Figure S1 (a)). Pooling reduces the number of parameters and induces some amount of translational invariance with respect to the position of objects in an input image. A crucial aspect for training neural networks is to use a non-linear activation function such as ReLU which preserves positive activations and sets negative activations to 0, i.e. $f(x) = \max(0, x)$. The transposed convolution (sometimes referred to as up-convolution, fractionally-strided convolution, or deconvolution) can be viewed as an inverse operation to max pooling which uses convolutional filters to up-sample pixels from a feature map to several pixels in a subsequent feature map. Finally, the dropout operation allows switching off individual neurons temporarily. For instance, with a dropout probability of 50% per neuron about half of the neurons can be switched off at random for a single training iteration. This usually improves the stability and accuracy of prediction outcomes.

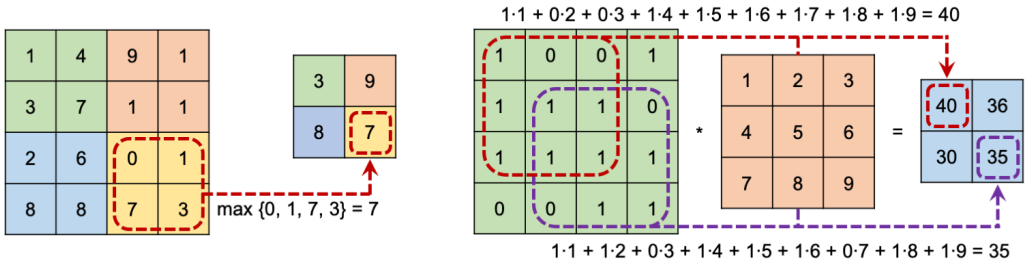

Figure S1. Illustration of the max pooling and convolution operation. In this study, for (a) max pooling, a kernel of size 2×2 was used with a stride of 2; i.e., the dashed box is shifted to the elements highlighted by the different colours. For (b) convolution, a kernel of size 3×3 (highlighted in orange) was used with a stride of 1; i.e., the dashed boxes are shifted by one column/row. The values in a patch of the input map (highlighted in green) are multiplied element-wise with the weights of the convolutional kernel, and the results are summed up, forming the feature map. For illustration purposes the weights are chosen from 1 to 9 and are subject to change during training.

S2. Mixed Thresholds for Trend Analysis

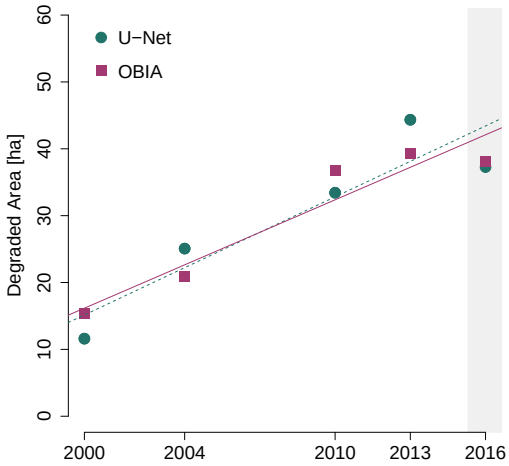

Figure S2. Total degraded area prediction of the U-Net with individually selected thresholds best suited for every erosion class on the held-out test region. The thresholds were selected according to a detailed threshold analysis (not shown) to be: 0.2 for shallow landslides, 0.2 for livestock trails, 0.3 for sheet erosion, 0.5 for management effects. Although deviations in the total degraded area persist, the linear trends of the two methods almost coincide.